Développement et application de méthodes pour l'analyse de la composition du microbiote humain dans un contexte clinique en utilisant des stratégies de séquençage alternatives

Benoît Goutorbe

université PARIS-SACLAY

Développement et application de méthodes pour l'analyse de la composition du microbiote humain dans un contexte clinique en utilisant des stratégies de séquençage alternatives

Development and application of methods to study human microbiota composition in a clinical context using alternative sequencing strategies

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, structure et dynamique des systèmes vivants
Spécialité de doctorat : Sciences de la vie et de la santé
Graduate School : Life Sciences and Health,
Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche
MaIAGE (Université Paris-Saclay, INRAE),
CRCM (INSERM, CNRS, Aix-Marseille Université, IPC),
et au sein du Laboratoire Alphabio,
sous la direction de Sophie SCHBATH, directrice de recherche,
la co-direction de Ghislain BIDAUT, ingénieur de recherche,
la co-direction de Philippe HALFON, professeur.

Thèse soutenue à Paris-Saclay, le 21 octobre 2022, par

Benoît GOUTORBE

Composition du jury

Membres du jury avec voix délibérative

Christophe AMBROISE Professeur, Université d'Évry Val d'Essonne	Président
Géraldine PASCAL Ingénieure de Recherche, HDR, INRAE centre Occitanie-Toulouse	Rapporteure & Examinatrice
Edi PRIFTI Chargé de Recherche, HDR, IRD, Sorbonne Université	Rapporteur & Examinateur
Emmanuelle LE CHATELIER Chargée de recherche, INRAE centre Île-de-France - Jouy-en-Josas - Antony	Examinatrice
Christophe MOUGEL Directeur de Recherche, INRAE centre Bretagne-Normandie	Examinateur
Hélène TOUZET Directrice de Recherche, CNRS, Université de Lille	Examinatrice

THESE DE DOCTORAT
NNT : 2022UPASL061

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Développement et application de méthodes pour l'analyse de la composition du microbiote humain dans un contexte clinique en utilisant des stratégies de séquençage alternatives
Mots clés : Microbiote ; Métagénomique ; NGS ; Bioinformatique ; Santé

Résumé : Le microbiote et ses liens avec la santé humaine sont aujourd'hui très étudiés. Le séquençage à haut débit a grandement contribué à l'essor fulgurant du domaine en permettant d'identifier et de quantifier finement tous les micro-organismes d'un échantillon donné, sans les contraintes préalables d'isolement et de culture. Deux stratégies de séquençage sont majoritairement utilisées : le métabarcoding, qui consiste à ne séquencer qu'un gène marqueur, ce qui est peu coûteux mais peu résolutif ; et la métagénomique *shotgun*, qui consiste à séquencer tout l'ADN présent dans un échantillon, permettant une meilleure résolution taxonomique et une caractérisation fonctionnelle, mais dont les coûts sont parfois prohibitifs. Cette thèse s'intéresse à deux stratégies de séquençage intermédiaires, plus résolutives que le métabarcoding tout en étant moins coûteuses que la métagénomique *shotgun*.
Nous avons dans un premier temps évalué la pertinence de la métagénomique *shotgun* à faible profondeur de séquençage pour l'étude du microbiote intestinal. Pour cela, nous avons développé un pipeline d'analyse fiable et optimisé, puis montré la capacité de notre méthode à reconstruire les profils taxonomiques et discriminer les patients. Dans un deuxième temps, nous avons développé une approche intégrative de métabarcoding multi-marqueurs pour caractériser le microbiote vaginal, qui permet de tirer parti de l'universalité et du pouvoir discriminant de chaque marqueur. Ce travail de thèse inclut également l'analyse de données issues de deux projets cliniques distincts, l'un portant sur la standardisation du protocole pré-analytique et l'autre sur les liens entre le microbiote intestinal et une maladie auto-immune, le lupus érythémateux systémique.

Title : Development and application of methods to study human microbiota composition in a clinical context using alternative sequencing strategies
Keywords : Microbiota ; Metagenomics ; NGS ; Bioinformatics ; Health

Abstract : The microbiota and its relation to human health are nowadays widely studied. High-throughput sequencing has greatly contributed to the rapid development of this field by allowing the identification and quantification of all microorganisms in a given sample, without the prior constraints of isolation and culture. Two sequencing strategies are mainly used : metabarcoding, which consists of sequencing only one marker gene, which is cost-efficient but not very precise ; and shotgun metagenomics, which consists of sequencing all the DNA present in a sample, allowing a better taxonomic resolution and a functional characterization but whose cost is sometimes prohibitive. This thesis focuses on two intermediate alternative sequencing strategies that are more precise than metabarcoding but less expensive than shotgun metagenomics.
We first evaluated the relevance of shallow shotgun metagenomics for studying the gut microbiota. We developed a reliable and optimized analysis pipeline, and then demonstrated the ability of this approach to reconstruct taxonomic profiles and differences between patients. In a second step, we developed an integrative multi-marker metabarcoding approach to characterize the vaginal microbiota, which takes advantage of the universality and discriminatory capacity of each marker. This thesis work also includes the analysis of data from two clinical projects, one on the standardization of the pre-analytical protocol and the other on the links between the gut microbiota and an autoimmune disease, systemic erythematosus lupus.

Remerciements

Je voudrais d'abord remercier les membres du jury, Christophe Ambroise le président, Géraldine Pascal et Edi Prifti les rapporteurs, Emmanuelle Le Chatelier et Chirstophe Mougel les examinateurs, d'avoir accepté d'évaluer mon travail de thèse.

Je remercie ma directrice de thèse Sophie Schbath, ainsi que mes co-directeurs Philippe Halfon et Ghislain Bidaut, pour m'avoir donné l'opportunité de réaliser cette thèse. La confiance que vous m'avez témoignée tout au long du projet m'a été précieuse. Merci à tous les trois pour la qualité de votre encadrement et la richesse de nos échanges qui m'ont permis de mener à bien ce projet.

Je remercie vivement mes encadrants à MaIAGE, Anne-Laure Abraham, Mahendra Mariadassou et Sophie Schbath. Merci de m'avoir guidé et soutenu pendant ces trois années. J'ai appris beaucoup à vos côtés, tant sur les aspects scientifiques que pédagogiques. Votre implication et votre bienveillance sans faille m'ont été indispensables. Je remercie également l'ensemble des personnes que j'ai pu côtoyer dans l'équipe StatInfOmics qui ont rendu mes séjours à Jouy agréables et enrichissants.

Je tiens à remercier mes collègues à Alphabio, particulièrement Anne Plauzolles, Marion Bonnet et Eya Toumi. La qualité de votre travail, notamment pour monter les projets, étudier la bibliographie, recruter les patients et mener à bien les manips, m'ont permis de travailler dans les meilleures conditions. Vos conseils m'ont été précieux à chaque étape de ma thèse. Je remercie également l'ensemble des membres, et anciens membres, de l'équipe microbiote avec qui j'ai eu le plaisir de collaborer : Aurélie, Sabrina, Lucie, Stella, Mohamed. Je remercie également les personnes que j'ai côtoyé à Alphabio : Guillaume, Vincent, Coralie, Sara, Phong, Cathy, France, et beaucoup d'autres. Vous avez tous contribué, par votre énergie, votre soutient et votre bonne humeur, à rendre ces années riches et agréables.

Je remercie également les membres du CRCM, et notamment de la plateforme Cibi, pour m'avoir accueilli dans l'équipe. Je joins à ces remerciements l'ensemble communauté bioinfo du CRCM et de l'IPC. Nos discussions dans l'*open space*, tant sur nos problèmes et astuces techniques que sur les aspects scientifiques de nos projets respectifs, m'ont beaucoup apporté, pendant mon stage puis pendant ma thèse. Je remercie notamment Julien, Quentin, Léonard, Adrien, Lucie, Samuel.

Je remercie l'ensemble des membres de la plateforme MIGALE pour la qualité des infrastructures de calcul et pour leur réactivité en cas de problème. Je remercie particulièrement Olivier Rué, pour son aide et ses conseils concernant la prise en main de FROGS, de DADA2, et sur l'utilisation de gyrB.

Enfin, je souhaite adresser un immense merci à mes proches pour m'avoir apporté tout le soutien dont j'ai eu besoin pendant ces années. Merci à mes parents et mes frères et sœur : notre unité, votre soutien inconditionnel, et votre volonté d'avancer m'ont porté. J'aurais été bien incapable de mener à bien ce projet sans vos encouragements. Je souhaite également remercier mes amis, l'ensemble de ma famille et toutes les personnes qui m'ont entouré dans les bons et les mauvais moments. Je remercie infiniment Vivi, qui m'a tant soutenu et apporté. Merci pour ça et pour tout le reste.

Table des matières

1 - INTRODUCTION

1.1 . Étude des microbes

Les premières observations des microbes ont été faites au $XVII^e$ siècle, quand Francesco Redi et Antoni van Leeuwenhoek ont observé grâce aux premiers microscopes ces formes de vie invisibles à l'oeil nu. Au XIX^e, on apprend à isoler et cultiver ces micro-organismes pour mieux les étudier. On comprend également qu'ils sont responsables d'infections. Ignas Semelweiss propose la théorie des germes, on découvre alors les concepts d'hygiène, et la vaccination se développe pour faire face aux épidémies. Dans la première moitié du XX^e siècle, Robert Koch propose ses postulats pour prouver la relation de causalité entre un pathogène et une maladie infectieuse et Alexander Flemming découvre la pénicilline, le premier antibiotique. On a compris dès l'époque de Louis Pasteur que toutes les bactéries n'étaient pas pathogènes et que certaines étaient impliquées dans les procédés de fermentation. Élie Metchnikoff, le découvreur de la phagocytose, avançait au début du XX^e siècle que certains micro-organismes contenus dans les yaourts pouvaient ralentir le vieillissement, proposant ainsi le fait que certains microbes pouvaient nous être bénéfiques. La seconde moitié du XX^e siècle a vu la naissance de la biologie moléculaire, à la suite de la découverte en 1952 de l'ADN comme support de l'information génétique grâce aux expériences de Alfred Hershey et Martha Chase [67], puis de la résolution de sa structure en double hélice par James Watson et Francis Crick en 1953 [186]. L'émergence de la biologie moléculaire a notamment bouleversé la façon de classifier les espèces bactériennes. Celles-ci étaient classées jusqu'alors de manière assez grossière, selon des critères morphologiques ainsi que sur certaines propriétés biochimiques et physiologiques, ou encore selon leur pathogénicité. Des critères moléculaires, comme un taux d'hybridation entre l'ADN de deux souches supérieur à 70%, ont été utilisés pour définir la notion d'espèce bactérienne [153]. Le séquençage de l'ADN a ensuite été permise par les travaux de Frederick Sanger en 1980, pour lesquels il reçut son deuxième prix Nobel. Le fait de pouvoir lire l'ADN a permis l'émergence de la phylogénie moléculaire, qui a de nouveau changé notre manière de classifier les bactéries, en s'appuyant sur des approches par similarité de séquences, d'abord sur un ou plusieurs gènes, puis sur tout le génome. Ces caractérisations moléculaires n'étaient cependant faites que sur des bactéries isolées et cultivées, du fait du faible rendement de la technique Sanger et de la nécessité d'avoir une concentration initiale d'ADN élevée pour pouvoir le séquencer. L'avènement dans le début des années 2000 du séquençage à haut débit, souvent appelé NGS pour *Next Generation Sequencing*, a levé cette limitation en permettant de lire plusieurs millions de séquences en parallèle, à partir de concentrations initiales d'ADN beaucoup plus faibles et en réduisant fortement le coût du séquençage. Les progrès réalisés à cette époque dans les protocoles d'extraction pour divers écosystèmes ont rendu possible l'analyse d'écosystèmes très divers. Ce sont les premiers pas de la métagénomique, également appelée génomique environnementale. Cette méthode a non seulement permis d'établir la composition des écosystèmes microbiens, en identifiant les micro-organismes et en estimant leurs abondances relatives, mais a aussi révélé la partie immergée de l'iceberg que constituent les bactéries non cultivables. On estime aujourd'hui que seuless 30% des espèces bactériennes retrouvées dans le tube digestif humain sont cultivées [5], et ce pourcentage est largement plus faible dans d'autres écosystèmes. Que ce soit dans le fond des océans, dans les sols ou dans nos intestins, on a alors pu véritablement se rendre compte de la diversité et de la complexité des écosystème microbiens et de leurs relations avec leur environnement. Deux stratégies de séquençage sont actuellement largement utilisées. D'une part le métabarcoding, descendant moderne des travaux pionniers de Carl Woese [187], consiste à ne séquencer qu'un gène marqueur dans le but d'identifier et quantifier les microbes

présents dans un écosystème. D'autre part la métagénomique *shotgun* consiste à séquencer tout l'ADN d'un écosystème, ce qui permet d'analyser non seulement la composition taxonomique mais aussi, après des étapes d'assemblage et d'annotation fonctionnelle, d'identifier les gènes et fonctions présents et de reconstruire les génomes des micro-organismes. Ces deux méthodes seront présentées en détail, en comparant leurs avantages et inconvénients, dans la section 1.3.

Le microbiote humain, défini comme l'ensemble des micro-organismes (bactéries, champignons et virus) colonisant les différentes parties du corps humain (appareils digestif et respiratoire, cavités buccale, nasale, vaginale, *etc.*), a pu être étudié à grande échelle grâce à ces développements. Les recherches sur le microbiote humain et ses liens avec la santé ont alors connu un essor fulgurant avec des projets exploratoires fondateurs : *MetaHIT* (*Metagenomics of Human Intestinal Tract*) en Europe [140] et le *HMP* (*Human Micobiome Project*) aux États-Unis [172]. Le projet *MetaHIT* a exploré le microbiote intestinal humain par métagénomique *shotgun*, ce qui a permis notamment de construire un catalogue de gènes retrouvés dans cet écosystème [140, 92]. Ce projet a également permis d'explorer la diversité du microbiote intestinal et de définir des compositions-type structurant la composition du microbiote des individus appelés entérotypes [11], même si le nombre et la nature exacte de ces derniers est controversée [32]. Il a également permis d'établir des corrélations entre la composition et la richesse du microbiote et le métabolisme de l'hôte [88]. Le *HMP* s'est intéressé au microbiote humain dans sa globalité (microbiotes intestinal, cutané, bucal, nasal et vaginal) en utilisant du métabarcoding, de la métagénomique *shotgun* et des approches par culture. Il a permis d'explorer la diversité des micro-organismes constituant ces écosystèmes et a contribué à construire des catalogues de séquences, de gènes et de génomes représentatifs du microbiote humain. À la suite de ces projets, il y a eu une explosion du nombre d'études s'intéressant aux liens entre le microbiote et la santé humaine. Elles ont permis de mettre en évidence l'implication du microbiote dans diverses fonctions de l'organisme, ainsi qu'une altération de la composition du microbiote dans de multiples pathologies.

La lecture de l'ADN par séquençage à haut-débit permet de caractériser un écosystème à un moment donné, en évaluant sa composition et éventuellement les fonctions métaboliques qu'apportent les différents micro-organismes qui le composent. Afin d'aller plus loin dans l'interprétation des résultats, et prendre en compte l'activité des micro-organismes, d'autres approches complémentaires peuvent être utilisées : la métatranscriptomique utilise le séquençage à haut-débit pour lire les ARN messagers (par opposition à l'ADN) renseignant ainsi sur l'activité des gènes ; la protéomique et la métabolomique, permettent d'analyser, par diverses techniques de spectrométrie, les protéines et métabolites fabriqués par les micro-organismes. Par ailleurs, la culture des micro-organismes à haut débit [85, 52, 199], appelée *culturomique*, est essentielle pour faire avancer les connaissances sur le domaine. La culturomique permet d'enrichir tant les biobanques que les catalogues de gènes et de génomes, qui sont utiles pour l'analyse des données de métagénomique (*cf.* section 1.3). Elle est également nécessaire pour l'étude détaillée des propriétés de certaines espèces d'intérêt. Réciproquement, la métagénomique peut servir à identifier les taxa importants qu'il serait intéressant de parvenir à cultiver, les *most wanted taxa* [51, 6], et définir des échantillons à partir desquels les isoler ainsi que les conditions dans lesquelles ils sont susceptibles d'être cultivés [124].

1.2 . Microbiote humain et santé

L'étude du microbiote humain et de ses liens avec la santé constituent aujourd'hui un champ de recherche à part entière. Les différentes révolutions technologiques présentées précédemment ont permis l'émergence de cette discipline, à l'interface entre la microbiologie, la médecine et les sciences computationnelles. Le but de ce paragraphe est d'illustrer pourquoi le microbiote humain suscite autant d'intérêt, mais aussi de souligner les perspectives et les verrous à franchir pour l'utilisation du microbiote humain dans un contexte clinique.

1.2.1 . L'homme symbiotique

Il est aujourd'hui admis que nous sommes des individus symbiotiques, composés approximativement d'autant de cellules humaines que de bactéries [158], et que les bactéries que nous abritons contiennent ~ 100 fois plus de gènes que notre propre génome [59]. Notre microbiote s'installe dès la naissance [146], voire même pendant la grossesse [31], et se trouve à chaque interface entre le corps humain et son environnement. Les microbiotes intestinal, respiratoire, buccal, vaginal, cutané, *etc.* sont autant d'écosystèmes évoluant dans des conditions physico-chimiques très différentes en interaction avec notre organisme. La symbiose désigne la cohabitation harmonieuse entre notre organisme et ces microbes, que nous abritons et nourrissons en échange des nombreuses fonctions qu'ils nous apportent.

C'est dans notre tube digestif que se situent le plus grand nombre et la plus forte diversité de microbes, qui constituent ce qu'on appelle aujourd'hui le microbiote intestinal. Le microbiote intestinal est impliqué, entre autres, dans la digestion des aliments, nous apportant des fonctions métaboliques essentielles. Par exemple, le microbiote permet de dégrader certaines fibres alimentaires (polysaccharides) en acides gras à chaînes courtes (AGCCs). Les AGCCs sont eux-mêmes au coeur de nombreuses fonctions : le butyrate est par exemple la source principale d'énergie des colonocytes (cellules épithéliales du colon) et joue ainsi un rôle essentiel pour le maintien de la barrière intestinale [113, 103], tandis que d'autres AGCCs sont impliqués dans le métabolisme des lipides [66] ou encore dans la régulation de la satiété [24]. Le microbiote intestinal joue également un rôle important dans la maturation du système immunitaire [57, 55]. Le mode d'accouchement, par voie basse ou par césarienne, via son influence sur l'acquisition précoce et la composition du microbiote humain dans les premiers mois de la vie [149], pourrait être impliqué dans certains troubles de l'immunité [171, 34, 177]. Le microbiote intestinal joue par ailleurs un rôle de barrière, qui nous aide à lutter contre la prolifération de pathogènes [74, 77].

Cet état symbiotique n'est pas complètement compris, bien que l'on ait élucidé certains mécanismes d'action contribuant à l'équilibre et l'inter-dépendance entre un individu et son microbiote. La description et la définition de ce qu'est un microbiote sain est aujourd'hui très partielle [159]. On sait que le microbiote intestinal doit être diversifié, qu'il doit est souvent dominé par certains phyla bactériens, que l'absence de certaines espèces dites fondatrices peut être néfaste, *etc.* Cependant, on observe une immense diversité dans la composition du microbiote des individus, même si ceux-ci sont en bonne santé. On a établi que de nombreux facteurs influencent la composition du microbiote humain, comme la situation géographique, l'ethnicité, l'âge, l'alimentation, la prise de médicaments, *etc.* [145, 74]. Cette grande diversité observée dans un microbiote équilibré, complique la compréhension de ce qu'est un microbiote déséquilibré.

1.2.2 . Le microbiote humain dans les pathologies

Dans de très nombreuses pathologies humaines, des altérations du microbiote ont été observées. Plusieurs paramètres de la symbiose décrite dans le paragraphe précédent peuvent être affectés, on parle de dysbiose. Un cercle vicieux peut s'installer, avec une modification de la composition du microbiote digestif, un affaiblissement de la perméabilité membranaire, et finalement une stimulation de la réaction immunitaire.

On a trouvé des altérations dans la composition du microbiote intestinal de patients atteints de maladies gastro-intestinales, et notamment les maladies inflammatoires de l'intestin (la colite ulcéreuse et la maladie de Crohn) [121, 131], mais aussi dans les maladies hépatiques [173, 95], les maladies métaboliques, notamment le diabète de type II [63] et l'obésité [94], ou encore les maladies auto-immunes [39]. L'axe intestin-cerveau est aujourd'hui très étudié, à travers notamment l'implication du microbiote dans les maladies neuro-psychiatriques comme la dépression [119, 106] ou les troubles du spectre de l'autisme [72, 30]. De nombreuses études ont également mis en évidence les liens entre le microbiote et certains cancers, le microbiote pouvant intervenir comme un facteur favorisant le développement de cancers colorectaux [56], mais également expliquer, voire moduler la réponse aux traitements anti-cancéreux [3, 197, 194].

1.2.3 . L'utilisation du microbiote humain dans un parcours de soin

Nous allons désormais nous intéresser à certains exemples montrant comment le microbiote et son analyse peuvent être utilisés dans un parcours de soin.

Un marqueur non-invasif Dans de nombreux cas, c'est le caractère non-invasif du prélèvement qui est particulièrement intéressant, l'analyse de la composition du microbiote se faisant souvent par prélèvement de selles. Dans les maladies hépatiques par exemple, de nombreuses études ont montré que la composition du microbiote permettait d'évaluer le degrés d'atteinte du foie [141, 95, 126, 163]. Le microbiote pourrait ainsi à terme être utilisé pour suivre la progression de la maladie chez un patient ayant une atteinte hépatique, et ajuster la prise en charge si une dégradation est observée.
Dans un autre registre, plusieurs études ont également mis en évidence des altérations du microbiote intestinal chez les patients atteints de cancers colorectaux, suggérant qu'une inflammation liée au microbiote puisse agir comme un facteur de risque favorisant la carcinogénèse, mais aussi influençant la réponse au traitement [151, 56, 188]. Ainsi, l'analyse du microbiote pourrait servir à terme de marqueur diagnostique ou pronostique, comme moyen de dépistage en amont des coloscopies, étant non-invasif et moins coûteux, pour évaluer les risques de cancers colorectaux chez les patients [181, 127]. Plus récemment, dans le cas du cancer du pancréas, il a été montré que le microbiote, combiné à d'autres marqueurs non-invasifs, améliorait grandement la détection précoce du cancer [79]. Ces cancers étant souvent diagnostiqués trop tard (plus de la moitié des cancers du pancréas sont diagnostiqués alors que le cancer est déjà métastasé), la possibilité de dépistage précoce pourrait être d'une grande utilité.

Une médecine personnalisée Un exemple classique de l'utilisation du microbiote est la prédiction de la réponse à l'immunothérapie chez les patients atteints de cancer [93, 105]. La prédiction de la réponse par l'analyse du microbiote peut servir dans ces cas à orienter le traitement des patients, et ainsi augmenter ses chances de guérison.
Dans le cadre du diabète de type II, une étude [196] a exploré une approche originale pour utiliser le microbiote au service de la prise en charge des patients. Cette étude a montré que

la composition du microbiote, associée à d'autres paramètres biologiques, pouvait expliquer la variabilité de la réponse glycémique à différents repas d'un individu à l'autre. Ils ont ensuite utilisé des algorithmes d'apprentissage automatique pour prédire la réponse glycémique à un repas donné pour chaque patient, et ainsi ouvert la voie à une nutrition personnalisée pour les patients diabétiques, améliorant la qualité de vie des patients.

Des voies thérapeutiques Le microbiote est modulable : on peut agir sur lui, par différents moyens, pour en modifier la composition et ainsi espérer l'utiliser pour en tirer un bénéfice, offrant des perspectives thérapeutiques nouvelles dans une démarche préventive ou curative. L'étude des liens de causalité entre l'altération du microbiote et les pathologies est un aspect fondamental. Dans certains cas, agir sur le microbiote, en parallèle à d'autres traitements, peut permettre d'améliorer la condition de santé.
On peut agir sur le microbiote intestinal par une intervention sur le régime alimentaire, par l'administration de probiotiques, de prébiotiques, d'antibiotiques ou par transplantation fécale. Il a par exemple été montré que la diversité microbienne était rétablie lors d'interventions diététiques visant à traiter l'obésité [33] pouvant contribuer au retour à la normale des paramètres métaboliques des patients. Il a été montré que l'administration de probiotiques avait un effet bénéfique sur la perte de poids [161, 200], et pouvait avoir un effet limité mais significatif sur les troubles de l'anxiété et de dépression [87], ainsi que pour l'amélioration des symptômes chez des patients atteints de maladies inflammatoires de l'intestin [28]. La transplantation fécale, qu'elle soit par voie orale (sous forme de capsule) ou rectale, est d'ores et déjà utilisée pour traiter des infections récalcitrantes à *C. difficile* [73], et connaît des résultats prometteurs pour traiter les maladies de Crohn [166], du Lupus [70], ou encore des colites ulcéreuse [112, 65]. En cancérologie, après avoir identifié que le microbiote était associé à la réponse à l'immunothérapie [165, 104], une étude clinique a récemment montré que certains patients n'ayant pas répondu à une immunothérapie, y répondait après avoir reçu une transplantation fécale [37]. Une étude a également exploré avec succès l'utilité d'une auto-transplantation fécale dans le but de rétablir le microbiote à son état initial après une chimiothérapie associée à des antibiotiques, lors du traitement de la leucémie myéloïde aiguë, permettant ainsi de limiter les complications [100].
L'ensemble de ces exemples montrent que nous sommes à un tournant de la compréhension et de l'utilisation du microbiote. Si la recherche a connu un essor fulgurant depuis un peu plus d'une dizaine d'années, des connaissances solides se sont lentement construites et ouvrent la voie vers l'utilisation du microbiote dans des pratiques médicales courantes.

1.3 . Analyse des écosystèmes microbiens par séquençage à haut débit

Nous allons dans cette section détailler les objectifs de l'analyse des écosystèmes microbiens par séquençage à haut débit. Nous introduirons les deux stratégies de séquençages dominantes, le métabarcoding et la métagénomique *shotgun*, puis expliquerons les raisons qui ont motivé notre recherche de stratégies de séquençage alternatives.

1.3.1 . Objectifs de l'analyse des écosystèmes microbiens

Le séquençage à haut débit a apporté une révolution dans le domaine de l'analyse des écosystèmes microbiens, car il a permis pour la première fois de décrire de manière exhaustive les organismes présents dans un échantillon donné, sans connaissance préalable de ceux-ci. La question biologique sous-jacente à l'analyse guide les choix méthodologiques, mais certains types d'analyse sont communs à beaucoup d'études. Il ne s'agit pas d'une revue exhaustive des analyses pouvant être menées, mais d'une introduction aux notions utiles à la compréhension du manuscrit.

Profils taxonomiques Pour décrire un écosystème, il est fondamental d'identifier et de quantifier les organismes qui le composent. L'étape d'identification se fait en comparant les séquences lues lors de l'expérience de séquençage, appelées *reads*, avec une base de données de référence qui contient les séquences (gènes ou génomes, selon la stratégie de séquençage et les choix d'analyse) d'organismes connus, et dont la taxonomie a été préalablement déterminée. L'espèce est le rang taxonomique de référence pour identifier les organismes, mais l'identification peut se faire à un rang taxonomique supérieur ou inférieur, selon les méthodes utilisées. Le séquençage haut-débit ne permet pas une quantification absolue des micro-organismes, c'est-à-dire qu'on ne peut pas par exemple exprimer les observations en nombre de copies d'un organisme par volume ou masse d'échantillon total, mais on peut obtenir une quantification relative, c'est-à-dire donner la proportion des différents organismes. Chaque étape du protocole peut introduire des biais : la conditions de prélèvement et de conservation des échantillons, l'extraction de l'ADN, la stratégie de séquençage (*cf.* sections 1.3.2 et 1.3.3) et enfin du traitement bioinformatique. Ainsi, l'estimation qui est faite des abondances relatives est imparfaite. Cependant, lorsque l'on compare plusieurs échantillons qui ont suivi le même protocole, ces biais sont identiques sur tous les échantillons et les différences observées entre les abondances relatives estimées sont alors interprétables.

La liste des organismes présents et leurs abondances relatives constituent le profil taxonomique d'un échantillon. Le profil taxonomique est à la fois de nature hiérarchique, car on peut lire le profil à différents niveaux taxonomiques, et de nature compositionnelle, car on a accès aux proportions des différents organismes dans l'écosystème. Une représentation intéressante de ces données est fournie par l'outil `krona` [128] qui fournit une visualisation interactive des profils taxonomiques, comme illustré sur la figure 1.1.

Établir un profil taxonomique revient à évaluer la composition d'un écosystème, composition qui sera ensuite utilisée dans diverses analyses. Un profil taxonomique présente l'avantage d'être facilement interprétable, en utilisant les connaissances de la microbiologie sur les organismes que l'on a identifiés. En revanche, on est limité par les connaissances sur la taxonomie des organismes, notamment pour les micro-organismes non cultivés.

Dans le cadre des différents projets présentés dans ma thèse, et qui se déroulent dans un

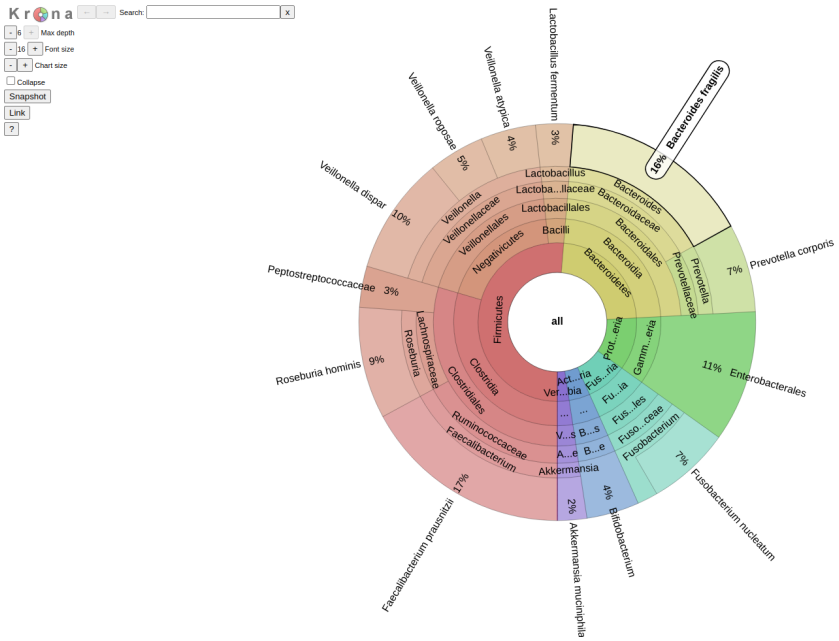


FIGURE 1.1 – Exemple de représentation d'un profil taxonomique d'une communauté microbienne par l'outils `krona`. La figure se lit de l'intérieur vers l'extérieur, représentant les différents niveaux taxonomiques : phylum, classe, ordre, famille, genre et espèce.

contexte clinique, obtenir des profils taxonomiques de bonne qualité, à la fois précis, avec une résolution taxonomique au moins jusqu'à l'espèce, et fiables, avec une maîtrise des faux positifs, est un objectif en soit, en plus d'être une donnée de base utilisée pour d'autres analyses.

Profils fonctionnels Au-delà de l'identification des organismes, il peut être intéressant d'identifier les gènes microbiens présents dans un écosystème afin d'avoir accès au potentiel fonctionnel du microbiote, et notamment les voies métaboliques présentes. Cela peut être fait en alignant les *reads* sur des gènes ou des génomes annotés. Il ne s'agit que d'un potentiel fonctionnel car on ne sait pas si ces gènes sont exprimés, à l'inverse de la métatranscriptomique qui séquence les ARN messagers. Le fait qu'un gène soit retrouvé en plus ou moins forte abondance dans un échantillon n'est donc pas forcément corrélé avec le rôle qu'il joue dans l'écosystème. Par ailleurs, les profils fonctionnels peuvent être difficilement interprétables car les fonctions associées aux différents gènes ne sont souvent que putatives, étant extrapolées par similarité de séquences avec des gènes dont on connaît la fonction. En revanche, ces données ont l'avantage de s'affranchir des limitations des connaissances taxonomiques des micro-organismes, et peuvent également servir de données de bases pour des analyses subséquentes.

α-diversité La notion d'α-diversité désigne la diversité au sein d'un échantillon. Cette notion est au coeur de nombreuses analyses d'écosystèmes : une α-diversité élevée dans le microbiote intestinal humain est généralement associée à une bonne santé alors que c'est l'inverse dans le microbiote vaginal. Il existe plusieurs mesures qui évaluent l'α-diversité selon différents critères. La richesse dénombre les entités taxonomiques (espèces, OTUs [1]) ou fonctionnelles (gènes, famille de gènes) distinctes détectées dans les échantillons. Les indices de Shannon ($-\sum_i p_i log_2 p_i$

1. *cf.* section 1.3.2

avec p_i l'abondance relative du taxa i) et de Simpson ($1 - \sum_i p_i^2$) tiennent compte des abondances relatives, et donnent une mesure d'α-diversité plus élevée si les abondances relatives des entités sont équitablement distribuées que si une entité domine toutes les autres. Enfin d'autres indices comme la mesure de Faith [45] prennent en compte la phylogénie des organismes présents. Quel que soit l'indicateur choisi, on obtient une mesure par échantillon, et on peut par exemple regarder s'il y a une différence d'α-diversité entre deux groupes de patients, via une analyse statistique de la variance (*anova*) ou un test de Wilcoxon-Mann-Whitney (appelé Wilcoxon dans le suite du manuscrit).

β-diversité La notion de β-diversité désigne la diversité, ou distance, entre plusieurs échantillons basée sur leurs profils taxonomiques. Il existe aussi différentes métriques pour mesurer cette diversité inter-échantillons : la distance de Jaccard ($1 - \frac{|(\Omega_A \cap \Omega_B)|}{|(\Omega_A \cup \Omega_B)|}$, avec Ω_A et Ω_B les ensembles de taxa identifiés dans les échantillons A et B respectivement) compare la composition de deux échantillons sur un critère de présence/absence des entités étudiées ; la distance de Bray-Curtis ($1 - 2\frac{\sum_i min(p_{i,A}, p_{i,B})}{\sum_i (p_{i,A} + p_{i,B})}$) tient compte des abondances relatives, tout comme la distance d'Aitchison (distance euclidienne après une transformation par *centered log ratio* des abondances relatives des taxa) [2] qui prend en compte le caractère compositionnel des données ; la distance Unifrac, et sa version pondérée pour prendre en compte les abondances relatives, utilisent la distance phylogénétique entre les taxa des deux échantillons. Une fois la matrice de distances entre paires d'échantillons calculée, on peut réaliser différents types d'analyse, comme par exemple une analyse en coordonnées principales (PCoA) pour réduire le nombre de dimensions des données, et ainsi projeter et visualiser les échantillons sur les axes principaux. On peut également regarder l'effet des métadonnées, c'est-à-dire des données associées à chaque échantillon (cas ou contrôle, âge, sexe, ou tout autre information pertinente pour la question biologique), sur les distances en réalisant une PERMANOVA.

Recherche de biomarqueurs Lorsque l'on compare deux groupes d'échantillons, on cherche souvent à savoir quelles bactéries sont différentiellement abondantes entre les deux groupes. Cette démarche peut-être appliquée à différentes entités taxonomiques, ou bien à des gènes. Cette étape est une démarche exploratoire, et les entités différentiellement abondantes que l'on identifie, appelées biomarqueurs, sont espérées être en lien avec la condition expérimentale à laquelle on s'intéresse. Il existe plusieurs méthodes pour réaliser ces tests. Les données étant non-gaussiennes, de grandes dimensions, avec une sur-abondance de zéros et compositionnelles, elles ne sont pas simples à manipuler. Le test non paramétrique de Wilcoxon-Mann-Whitney est souvent utilisé car il ne nécessite pas de condition particulière. Il n'est cependant pas le test le plus puissant, et des méthodes dédiées peuvent être plus performantes. Des méthodes comme `DESeq2` et `edgeR`, qui ont été initialement développées pour traiter des données de transcriptomique sont très souvent utilisées car ces données, provenant également de séquençage à haut débit, sont assez similaires à celles de données métagénomiques. Il existe également des méthodes comme `ALDEx2` [49] et `ANCOM` [101], qui exploitent l'aspect compositionnel des données en utilisant des transformations dédiées telles que les transformations *clr* (pour *centered log ratio*) ou *alr* (pour *additive log ratio*) respectivement.

Classification de patients Dans certaines études, il est pertinent de réaliser une classification des patients, qui peut être supervisée ou non. Les classifications supervisées correspondent à des cas où l'on cherche à utiliser le pouvoir prédictif du microbiote, dans une démarche diagnostique

ou pronostique. Dans ces cas, on utilise des données pour entraîner un modèle de classification, comme les forêts aléatoires, les machines à vecteurs de support (SVM) ou encore les réseaux de neurones [89]. Ces modèles peuvent être très puissants mais nécessitent de grandes cohortes d'entraînement et surtout des cohortes de validation pour s'assurer de la qualité des prédictions sur des données indépendantes du jeu de données d'entraînement. Dans certaines situations, on peut également réaliser une classification non-supervisée, comme la classification ascendante hiérarchique, pour identifier des sous-groupes de patients en se basant sur la composition de leur microbiote.

Pour réaliser toutes ces analyses, deux stratégies de séquençage sont majoritairement utilisées : le métabarcoding et la métagénomique *shotgun*. Nous allons désormais détailler leur principe de fonctionnement, leurs avantages et inconvénients.

1.3.2 . Métabarcoding

Le métabarcoding, également appelé métagénomique ciblée ou métagénomique amplicon ou encore métataxonomique, est la méthode la plus utilisée pour caractériser les écosystèmes microbiens. Cette stratégie de séquençage repose sur l'amplification par PCR d'un gène marqueur, ou d'une région d'un gène marqueur, censé permettre l'identification taxonomique des organismes. Le terme *métabarcoding* désigne d'ailleurs très bien l'idée de code barre que l'on peut lire pour identifier l'organisme.

Principe de fonctionnement Après avoir collecté un échantillon, l'ADN en est extrait, puis le gène ciblé est amplifié par PCR (*Polymerase Chain Reaction*) avant d'être séquencé. L'amplification par PCR repose sur la complémentarité entre une paire d'amorces (séquences d'une vingtaine de nucléotides) et la région ciblée. Par le contrôle de la température, on peut dénaturer l'ADN (séparer les deux brins), puis permettre aux amorces de s'hybrider à l'ADN, et enfin laisser à l'ADN polymérase le soin de synthétiser l'ADN contenu entre la paire d'amorces. On peut ensuite répéter ce cycle, et ainsi amplifier la région souhaitée. Le choix de la cible et la conception des amorces sont essentiels. On ajoute ensuite des index au séquences permettant d'identifier les différents échantillons qui peuvent alors être séquencés lors d'un même *run* de séquençage. Le séquençage en lui même peut alors être réalisé. La technologie Illumina est la plus utilisée à l'heure actuelle, et produit des *reads* appariés d'une longueur allant jusqu'à 2*300 nucléotides selon les appareils.

Choix du gène marqueur Pour mener à bien une expérience de *métabarcoding*, il faut choisir un marqueur, généralement un gène ou une région d'un gène. Dans un scénario idéal, ce marqueur devrait satisfaire les critères suivants :

- être présent et amplifié chez tous les organismes d'intérêts, ce qui stipule que ses extrémités soient conservées pour permettre la conception d'amorces de PCR universelles ;
- sa diversité nucléotidique doit permettre de discriminer toutes les espèces, voire sous-espèces ;
- les bases de données de ce gène marqueur doivent contenir tous les organismes qui sont présents dans l'écosystème étudié ;
- être présent en une seule copie dans les génomes ;
- ne pas être sujet aux transferts horizontaux entre les micro-organismes.

Pour l'étude des bactéries, qui constituent la grande majorité des micro-organismes du microbiote humain, le gène codant pour l'ARN 16S (petite sous-unité du ribosome) est utilisé

pour l'immense majorité des études. Il existe cependant d'autres cibles alternatives pour les bactéries, qui seront détaillées dans la section 3.1.1. Le gène codant pour l'ARN 16S est présent chez toutes les bactéries et les archées, et possède une structure très particulière, avec une succession de régions conservées permettant de concevoir des amorces universelles et de régions hypervariables permettant de discriminer les bactéries, comme illustré dans la figure 1.2, ce qui en fait un excellent candidat pour le métabarcoding. Il n'est cependant pas parfait, et les profils taxonomiques résultants souffrent de biais, qui seront détaillés dans la table 1.1.

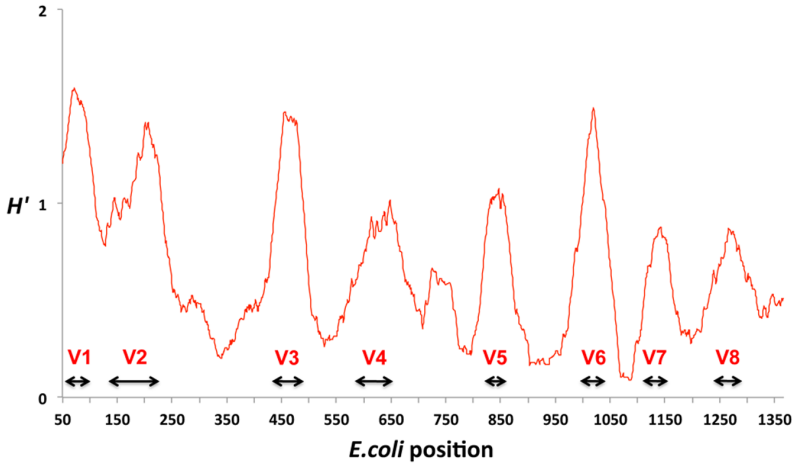


FIGURE 1.2 – Représentation de la variabilité le long du gène 16S et désignation des régions hypervaribales, issue de Vasileiadis *et al.*[180] : mesure d'entropie par position.

Analyses des données de métabarcoding Un pipeline pour l'analyse de données de métabarcoding comprend classiquement une première étape de filtre qualité, pendant lequel on retire les *reads* jugés de mauvaise qualité et les amorces de séquençage. Selon la taille de la région amplifiée et la taille des *reads*, il peut y a avoir un chevauchement entre les *reads forward* et *reverse*. Si c'est le cas, on utilise souvent ce chevauchement pour procéder à l'assemblage des paires de *reads*. Il y a ensuite une étape qui consiste à débruiter les données, soit en les agglomérant par similarité, on parle alors d'OTUs (pour *Operational Taxonomic Units* [26], soit en ajustant un modèle d'erreurs qui permet de détecter les variants biologiques des variants créés par les erreurs de séquençage, on parle alors d'ASVs [25]. La construction d'OTUs regroupe les séquences similaires qui correspondent parfois à plusieurs copies du marqueur au sein du même organisme ou d'organismes très proches, réduisant ainsi la dimensionnalité des données. L'approche par ASVs conserve toute la diversité biologique des séquences, ce qui est utile si les séquences provenant de plusieurs organismes d'intérêt sont très proches [25] mais peut surestimer l'α-diversité [75, 139]. Les OTUs ou ASVs rares ainsi que les chimères introduites par la PCR sont retirées. On peut ensuite réaliser une assignation taxonomique, en alignant les données sur une base de données de référence [98, 147], ou en utilisant une approche par classification [185, 198], ce qui permet de construire le profil taxonomique. On peut également réaliser un alignement multiple entre les séquences d'un jeu de données [80], ce qui permet de construire un arbre phylogénétique [138] des séquences du jeu de données, qui approxime l'arbre phylogénétique des organismes de l'écosystème. Il existe également des outils qui utilisent les profils taxonomiques pour inférer des profils fonctionnels, en utilisant les génomes de référence des espèces identifiées en métabarcoding [12, 42] mais la qualité de la prédiction est limitée

[170]
Grâce à l'utilisation massive du métabarcoding, il existe de nombreuses ressources à disposition de la communauté. Des pipelines faciles d'utilisation permettent de réaliser l'ensemble de l'analyse sans besoin d'infrastructure de calcul spécifique ni de compétences avancées en bioinformatique, par exemple FROGS, Mothur ou QIIME2 [44, 154, 18]. Cette technique profite également d'avoir des bases de données extrêmement fournies pour les marqueurs les plus utilisés, comme Silva [142], greengenes [176] et RDP [29] pour le 16S ou Unite [120] pour l'ITS, contenant des séquences représentatives de microbes provenant d'écosystèmes très divers.

Avantages et inconvénients du métabarcoding Au-delà de son faible coût, le métabarcoding bénéficie de nombreux outils d'analyse faciles d'utilisation, le rendant simple à mettre en œuvre. Les données brutes étant relativement légères (~ 50 Mo/échantillon), la technique est applicable pour des études contenant beaucoup d'échantillons. Il permet également d'obtenir une assignation taxonomique même si elle n'est pas toujours précise au niveau de l'espèce et de calculer des mesures d'α-diversité et de β-diversité. Par ailleurs, cette stratégie est applicable pour tout type d'écosystèmes, même en cas de faible quantité et qualité d'ADN, ou en cas de forte contamination par l'ADN de l'hôte par exemple.

Le marqueur parfait pour faire du métabarcoding n'existe pas, et les résultats obtenus par métabarcoding présentent ainsi des biais, représentés dans la table 1.1, qui peuvent être dus (i) au manque d'universalité du gène marqueur et des amorces PCR pour amplifier le marqueur, (ii) au nombre de copies du marqueur dans le génome, (iii) au faible pouvoir discriminant entre espèces et (iv) au manque de complétude des bases de données associées. Premièrement, les marqueurs sont souvent spécifiques à un domaine du vivant, par exemple l'ARN 16S n'est présent que chez les bactéries et les archées, tandis que les amorces ciblant l'ITS sont souvent spécifiques au microbiote fongique, et les virus ne sont pas ciblés en métabarcoding. Même au sein d'un domaine, les organismes dont le gène marqueur n'est pas ou peu amplifié (absence du gène, amorces n'amplifiant pas ou peu le gène dus à des *mismatchs* entre les amorces et le génome) seront absents ou sous-représentés dans les profils taxonomiques résultants. De plus, si le nombre de copies du marqueur dans le génome varie chez les différents organismes constituant l'écosystème, l'estimation de leurs abondances relatives sera erronée [19, 96]. Enfin, si la diversité nucléotidique au sein de la région amplifiée ne permet pas de discriminer toutes les espèces entre elles ou si la base de données de référence ne contient pas de séquence provenant d'organismes suffisamment proches de ceux retrouvés dans l'écosystème, ces organismes auront une assignation taxonomique à un niveau taxonomique supérieur à celui souhaité (genre, famille, *etc.* au lieu de l'espèce).

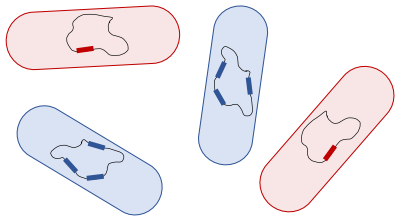
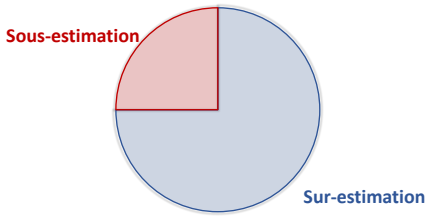
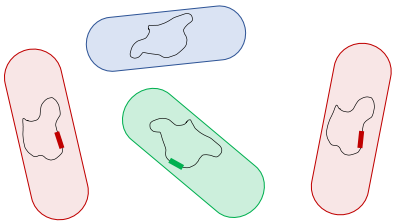
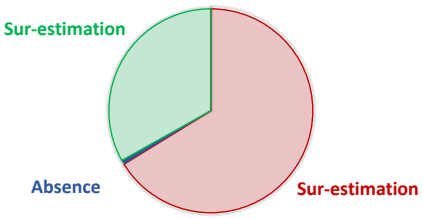
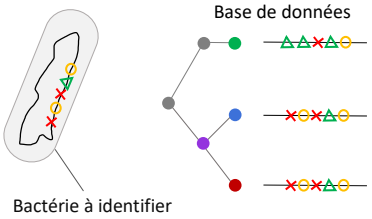
	Situation	Profil taxonomique résultant
Biais du nombre de copies		
Biais d'universalité du marqueur ou des amorces de PCR		
Limite du pouvoir discriminant		

TABLE 1.1 – Biais et limitations du métabarcoding

1.3.3 . Métagénomique *shotgun*

Principe de la métagénomique *shotgun* Cette stratégie de séquençage vise à séquencer l'ensemble du matériel ADN présent dans un échantillon. Pour cela, après collecte des échantillons et extraction de l'ADN, une étape de fragmentation aléatoire de l'ADN est réalisée. Les fragments d'ADN sont alors séquencés puis analysés.

Analyse des données de métagénomique *shotgun* Il existe plusieurs stratégies d'analyse pour les données de métagénomique *shotgun*, illustrées dans la figure 1.3, et le choix entre celles-ci dépend bien sûr des questions biologiques sous-jacentes au projet, de la profondeur de séquençage, c'est-à-dire du nombre de *reads* séquencés par échantillon, et des ressources à disposition. On peut distinguer deux grandes familles d'analyses : celles qui se basent sur un assemblage des *reads*, et celles qui se basent sur un alignement contre une base de données de

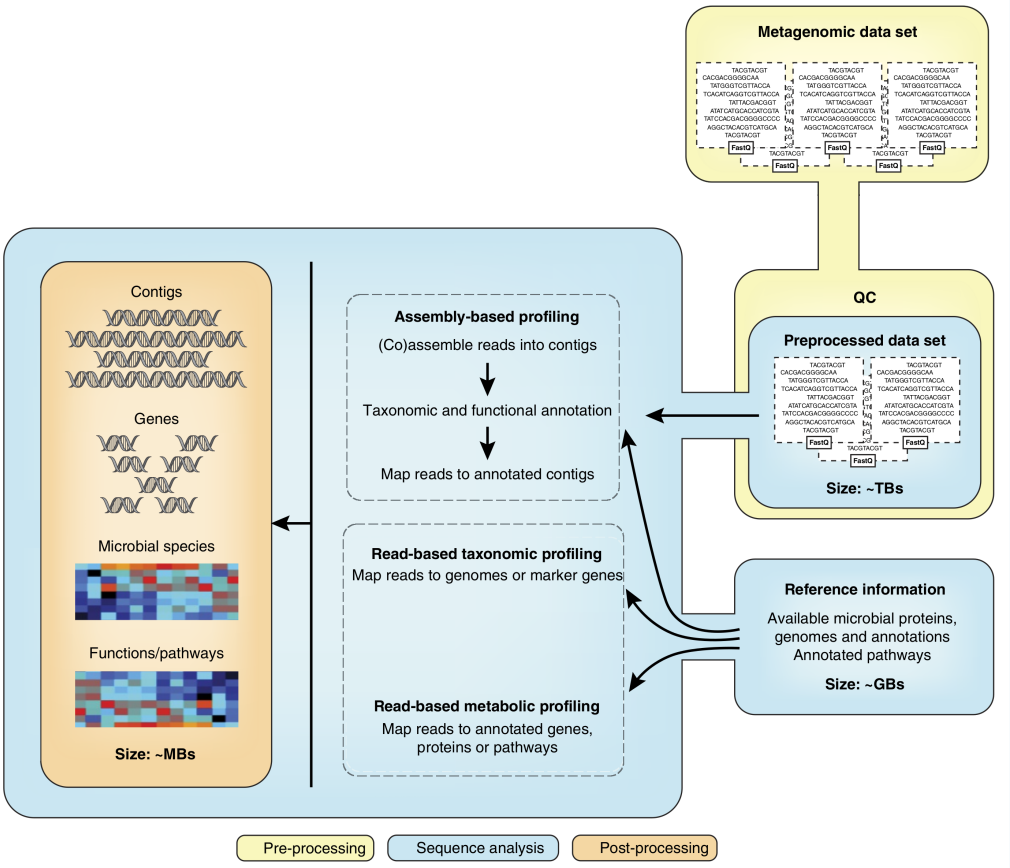


FIGURE 1.3 – Schéma adapté de Quince *et al.* 2017 [143]. Méthodes classiques pour analyser des données de métagénomique *shotgun* : après un contrôle qualité des *reads* (taille, taux d'erreurs, contamination de l'ADN de l'hôte) les données sont soit assemblées soit analysées par alignement des *reads* sur une base de données pour obtenir un profil taxonomique et/ou fonctionnel.

référence.

Méthodes basées sur un assemblage L'assemblage de données métagénomiques [21, 58] est une tâche extrêmement complexe, qui s'apparente à la reconstitution de plusieurs puzzles dont on aurait mélangé les pièces, avec des pièces manquantes, ne sachant ni le nombre total de puzzles, ni leur taille (certains étant plus grands que d'autres), ni les motifs contenus dans chaque puzzle. Les méthodes qui réalisent cette tâche utilisent le chevauchement entre les *reads* pour former des *contigs*, zones contiguës (plus longues que les *reads*) que l'on parvient à reconstituer. Ces méthodes ont recourt à des représentations des données sous forme de graphes de de Bruijn reliant chaque kmer présent dans un même *read* [122, 90]. On cherche ensuite souvent à regrouper les *contigs* que l'on suppose appartenir à la même espèce, lors d'une étape appelée le *binning*, en se basant sur le contenu nucléotidique des *contigs*, sur leurs co-occurrences dans les échantillons, et/ou en utilisant des sources de données externes (par exemple une annotation taxonomique des contigs), dans le but de produire des génomes [78, 7]. Ces génomes, qui sont souvent incomplets et peuvent contenir des contaminations [130], sont appelés MAGs, pour *metagenome-based assembled genomes*. Par ailleurs, il est possible d'explorer le potentiel fonctionnel des écosystèmes en réalisant une annotation structurale [156] et fonctionnelle des gènes

[71, 76]. Une approche alternative est de réaliser l'annotation structurale et fonctionnelle des gènes à partir des contigs, et de créer des catalogues de gènes. Les catalogues de gènes résultant de cette analyse peuvent être partitionnés en groupes de gènes co-occurrents et regroupés sous le terme de *metagenomic species* (MGSs) et *metagenomic Species Pan-genomes* (MSPs) [118, 135]. Ces analyses permettent d'explorer les écosystèmes microbiens sans connaissance préalable sur les organismes qui les composent. Dans le cas du microbiote humain, de grands projets [172, 92] ont permis de mettre en évidence la richesse, jusque là inexplorée, de la fraction non-cultivée du microbiote humain. Des études plus récentes utilisent les très grandes masses de données générées ces dernières années par différents projets, pour enrichir les catalogues de génomes et de gènes existants [4, 117, 132].

Ces analyses, extrêmement coûteuses en termes d'effort de séquençage et de traitement des données, ont permis néanmoins de produire des ressources qui peuvent aujourd'hui être utilisées pour l'analyse de nouvelles données.

Méthodes basées sur un alignement Grâce aux ressources à disposition pour la communauté, il est désormais possible d'analyser des données de métagénomique *shotgun* du microbiote intestinal humain par exemple, en alignant directement les *reads* séquencés sur les catalogues de gènes ou de génomes. Cela permet de réduire radicalement les ressources informatiques nécessaires au traitement des données, ainsi que la profondeur de séquençage nécessaire pour analyser les taxons rares. Il existe au moins 3 principales stratégies pour l'alignement de données métagénomiques : (1) l'alignement rapide sur des génomes complets à base de transformation de Burrows-Wheeler [91, 86] ou de *kmers* [189], (2) l'alignement sur un catalogue de gènes marqueurs des espèces [109, 157, 13], ou encore (3) l'alignement sur un catalogue de gènes ou de protéines [108, 92]. Dans ce dernier cas, les gènes du catalogue peuvent être assignées taxonomiquement à une espèce, permettant alors de construire un profil taxonomique, et ils peuvent également être associés à une fonction métabolique, permettant de construire un profil fonctionnel.

Avantages et inconvénients de la métagénomique *shotgun* La métagénomique *shotgun* présente de nombreux avantages. Elle permet d'analyser l'intégralité des écosystèmes microbiens, tant les bactéries et les archées, que les champignons ou les virus. Elle permet d'avoir une résolution taxonomique fine, jusqu'au niveau des espèces, voire des souches. Elle permet également d'explorer les capacités fonctionnelles du microbiote. En revanche, l'analyse du microbiote par métagénomique *shotgun* nécessite de générer beaucoup de données, et elle est donc plus coûteuse (*cf.* section 1.3.4). Les données sont souvent complexes à analyser, et nécessitent des infrastructures et des compétences adaptées.
Le taux de contamination par l'ADN de l'hôte peut aussi être un problème pour certains types de prélèvements. Contrairement au microbiote intestinal, les microbiotes vaginal, cutané et salivaires sont fortement contaminés par l'ADN humain lorsqu'ils sont analysés par métagénomique *shotgun*, comme montré dans la figure 1.4. Les *reads* humains sont écartés dès le pré-traitement des données lorsqu'on s'intéresse au microbiote humain. Le taux d'ADN humain peut dépasser les 90% dans certains prélèvements, ce qui rend le séquençage extrêmement coûteux si l'on souhaite obtenir suffisamment de *reads* bactériens. Il existe des méthodes biochimiques pour limiter le taux d'ADN humain avant le séquençage, mais elles introduisent un biais supplémentaire sur les estimations des abondances relatives [1, 160].

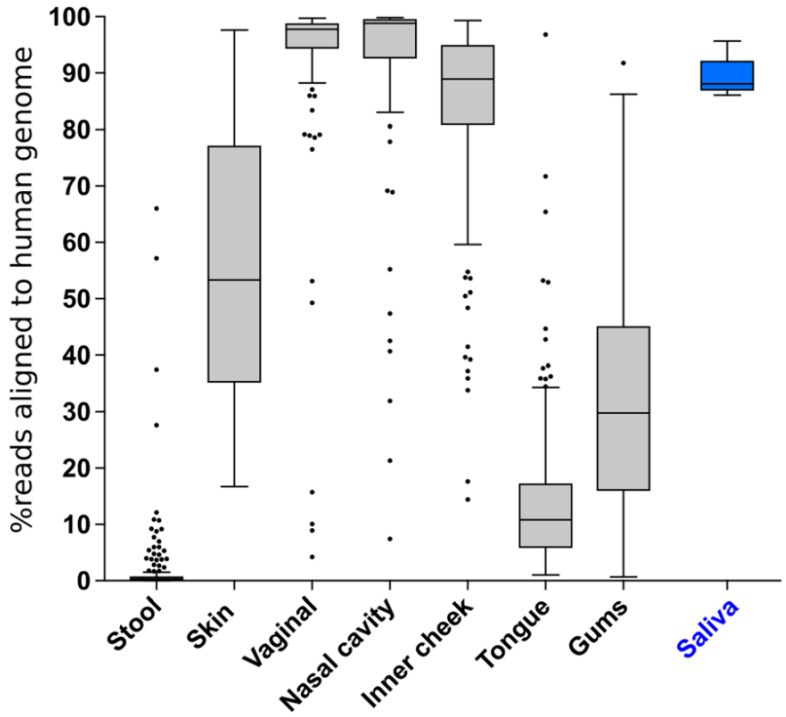


FIGURE 1.4 – Taux d'ADN humain dans les données de métagénomique *shotgun* par type de prélèvement. Figure issue de Marotz *et al.*[102] reprenant les données de HMP [172].

1.3.4 . Intérêt de stratégies alternatives

On a vu dans les sections 1.3.2 et 1.3.3 que chacune des méthodes possède des avantages et des inconvénients. Le métabarcoding présente l'avantage d'être peu coûteux, très facile d'utilisation, et de convenir à tout-types d'échantillons (notamment ceux fortement contaminés par l'ADN de l'hôte). En revanche, cette technique est peu résolutive (on ne peut pas tout le temps identifier les espèces), il est limité à un domaine (les bactéries et archées pour le 16S), et souffre de biais liés à l'universalité des amorces et aux nombres de copies du gène marqueur utilisé. La métagénomique *shotgun* permet une meilleure résolution taxonomique, qui ne se limite pas à un domaine et qui n'est pas biaisée par une amplification PCR, ainsi qu'une caractérisation fonctionnelle des écosystèmes. Cependant, cette technique est plus coûteuse et les données peuvent être plus compliquées à analyser et interpréter. Dans le cadre du microbiote intestinal, la métagénomique *shotgun* est souvent plébiscitée, pour sa meilleure résolution taxonomique et la possibilité d'étudier les profils fonctionnels [43, 87, 15, 82].

Pour étudier les liens entre le microbiote intestinal et la santé, compte-tenu de la grande variabilité inter-individuelle, il est primordial d'analyser un grand nombre d'échantillons. Que ce soit pour inclure un plus grand nombre de patients, intégrer un suivi longitudinal (plusieurs prélèvements sur un même patient à différents points de temps), ou encore intégrer une cohorte de validation à un projet exploratoire, il est toujours intéressant d'analyser plus d'échantillons. Cela permet d'augmenter la puissance statistique, ainsi que la fiabilité et la reproductibilité des résultats. L'écart de coûts de séquençage est très important entre le métabarcoding ($\sim$ 20€ pour 50K paires de *reads* en utilisant la technologie Illumina *MiSeq*) et la métagénomique *shotgun* ($\sim$ 120€ pour 20M paires de *reads* en utilisant la technologie Illumina *NextSeq*). Cet écart de coût incite les chercheurs à conduire beaucoup de leurs projets en *métabarcoding*, favorisant,

à budget équivalent, le nombre d'échantillons analysés. On peut voir sur la figure 1.5 que le nombre d'articles faisant référence au métabarcoding dans l'analyse du microbiote intestinal humain aujourd'hui est encore largement plus important que le nombre d'articles mentionnant la métagénomique *shotgun*. Je n'ai pas vérifié manuellement les milliers de résultats de ces deux requêtes, donc cette figure doit être interprétée avec précaution. Elle témoigne néanmoins de la réalité du grand nombre d'études qui sont encore aujourd'hui menées en métabarcoding dans le domaine.

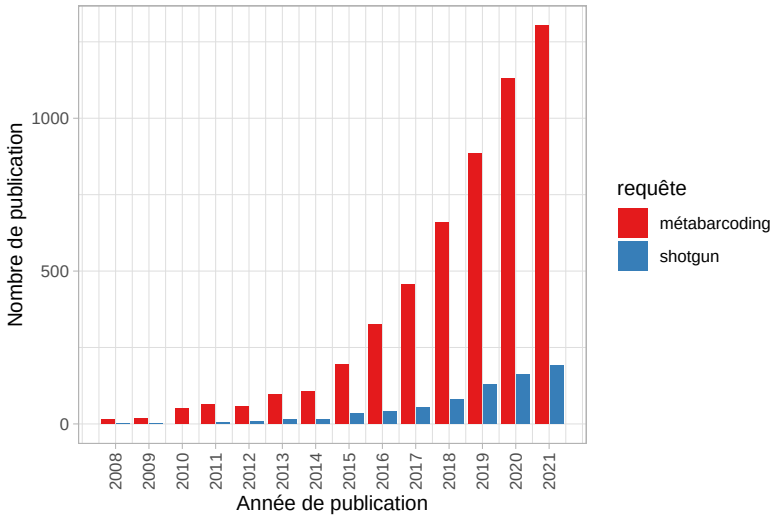


FIGURE 1.5 – Nombre d'articles référencés dans PubMed par année de publication avec les requêtes "*(human) AND (Gut microbiota) AND ((metabarcoding) OR (amplicon) OR (metataxonomics) OR (16S sequencing))*", notée *métabarcoding* et apparaissant en rouge, et "*(human) AND(Gut microbiota) AND ((shotgun metagenomics) OR (whole metagenome sequencing))*", notée *shotgun* et apparaissant en bleu.

Pour caractériser les microbiotes dont les échantillons sont fortement contaminés par l'ADN humain, notamment le microbiote vaginal, l'utilisation de la métagénomique *shotgun* est peu appropriée. Par ailleurs, comme nous le verrons dans la section 3.1.4, la faible résolution des profils taxonomiques obtenus par métabarcoding est problématique.
Ainsi, des stratégies de séquençage alternatives, plus résolutives que le métabarcoding, moins coûteuses que la métagénomique *shotgun*, et insensibles aux contaminations de l'ADN de l'hôte pourraient être utiles pour l'analyse du microbiote humain.

1.4 . Objectifs de la thèse

Cette thèse vise à développer et appliquer des méthodes d'analyse de la composition du microbiote en utilisant des stratégies de séquençage alternatives au métabarcoding et à la métagénomique *shotgun*. L'évaluation de la pertinence de ces stratégies de séquençage dans un contexte clinique est primordiale, tant pour la caractérisation fiable de la composition des écosystèmes que pour la capacité à discriminer les groupes de patients entre eux selon des caractéristiques cliniques d'intérêt.
Cette thèse a été menée en collaboration avec le Laboratoire Alphabio, un laboratoire de biolo-

gie médicale qui possède un service de Recherche et Développement s'intéressant au lien entre les microbiotes humains et la santé. Le projet d'Alphabio inclut une activité de recherche clinique, en collaboration avec l'Hôpital Européen de Marseille, ainsi qu'un service d'analyse de microbiote de routine. De par sa nature de laboratoire de biologie médicale, Alphabio s'intéresse particulièrement aux contextes cliniques dans lesquels l'analyse du microbiote peut être utile aux médecins pour améliorer le diagnostic et la prise en charge des patients. Dans le cadre de cette thèse, nous nous sommes concentrés sur la caractérisation des microbiotes intestinal et vaginal, pour lesquels les problématiques sont très différentes. Dans les deux cas, Alphabio a jusqu'alors utilisé le métabarcoding 16S, et ce projet de thèse s'est construit dans le but d'étudier les stratégies de séquençage alternatives permettant d'améliorer la résolution des profils taxonomiques obtenus.
Le chapitre 2 explore la métagénomiqe *shotgun* à faible profondeur, dite *shallow shotgun metagenomics* (SSM), pour l'analyse du microbiote intestinal. Cette approche a été proposée récemment par Hillman *et al.* [69] et consiste en une réduction radicale de la profondeur de séquençage par rapport à ce qui est classiquement utilisé, diminuant ainsi les coûts subséquents. A condition de disposer d'un catalogue de gènes et/ou génomes de référence pour l'écosystème étudié, la perte d'information liée à la réduction de séquençage est modérée. Le projet a pour but de (1) développer un pipeline bioinformatique pour construire des profils taxonomiques fiables et précis à partir des données de SSM, et de caractériser l'impact de la profondeur de séquençage sur (2) les profils taxonomiques obtenus et (3) la discrimination de patients selon des conditions cliniques d'intérêt.

Le chapitre 3 s'intéresse au métabarcoding *multi-marqueurs* pour la caractérisation du microbiote vaginal. Si la complémentarité entre différents marqueurs est établie dans la littérature, il n'existait pas de méthode permettant d'intégrer les données issues de plusieurs marqueurs pour produire un profil taxonomique consensus. L'objectif de ce projet est de proposer une méthode permettant de tirer parti de la complémentarité entre les marqueurs pour améliorer l'assignation taxonomique et l'estimation des abondances relatives. Nous étudierons ensuite la pertinence de cette approche pour l'analyse du microbiote vaginal.

Enfin, ce travail de thèse inclut également l'analyse de données issues de deux projets menés au sein d'Alphabio utilisant le métabarcoding. Le premier projet s'intéressait à la standardisation des protocoles pré-analytiques pour l'analyse du microbiote intestinal. L'objectif était d'étudier les biais liés aux conditions de conservation des échantillons avant leur analyse, et de déterminer les conditions qui minimisaient ces biais. Le deuxième projet portait sur les liens entre une maladie auto-immune, le lupus érythémateux systémique (SLE), et le microbiote intestinal. Le but de ce projet était de caractériser l'altération du microbiote chez les patients atteints de SLE, d'étudier les liens potentiels avec l'activité de la maladie, et de proposer une signature universelle cohérente avec les observations sur un modèle murin, et avec les données de la littérature. Ces projets ont fait l'objet de deux publications présentées au chapitre 4.

2 - MÉTAGÉNOMIQUE SHOTGUN A FAIBLE PROFONDEUR

2.1 . Contexte

Dans ce chapitre, je vais vous présenter l'intérêt et les limites de l'utilisation de la métagénomique *shotgun* à faible profondeur (notée SSM pour *shallow shotgun metagenomics* dans la suite du chapitre) pour l'analyse du microbiote intestinal humain dans des applications cliniques. Cette technique est une variante de la métagénomique *shotgun* "classique", et consiste à ne séquencer les échantillons qu'à très faible profondeur (*i.e.* peu de *reads* par échantillon). Pour étudier un écosystème aussi complexe que le microbiote intestinal, une profondeur de plusieurs dizaines de millions de *reads* par échantillon est classiquement utilisée alors que l'approche SSM nécessite typiquement ~ 1 millions de *reads* par échantillon. Comme nous le verrons, la SSM entraîne une réduction drastique de la quantité de données générées, et des coûts afférents, et nécessite des méthodes de traitement adaptées afin de perdre le moins d'information possible. Nous allons dans cette section introductive rappeler les outils d'analyse de données de métagénomique *shotgun* applicable à la SSM (section 2.1.1) puis faire un état de l'art des résultats obtenus sur l'impact de la profondeur de séquençage (section 2.1.2), avant de préciser les objectifs de mon travail (section 2.1.3).

2.1.1 . Méthodes d'analyse de données *shotgun*

Comme décrit dans la section 1.3.3, il existe plusieurs stratégies d'analyse pour les données de métagénomique *shotgun*. Dans le cadre de la SSM, la faible profondeur de séquençage rend l'assemblage totalement impossible. Par conséquent, cette approche n'est possible que si l'on dispose d'un catalogue de référence représentatif de l'écosystème que l'on étudie sur lequel aligner les *reads* de séquençage. Pour l'étude du microbiote intestinal humain, les catalogues dont on dispose actuellement [92, 5] sont issus de plus d'une décennie d'efforts de séquençage et contiennent à la fois des bactéries cultivées et non-cultivées, permettant d'explorer l'ensemble de l'écosystème.

Parmi les différentes approches par alignement, nous nous sommes tournés vers un alignement sur génomes complets, dans le soucis d'utiliser un maximum de *reads* pour construire le profil taxonomique, et dans un contexte où la profondeur de séquençage sera critique. L'approche par alignement sur des gènes marqueurs spécifiques à certaines clades n'utilise qu'une fraction des *reads*, ce qui est sous-optimal dans le cadre de la SSM. L'alignement sur un catalogue de gènes (incluant des gènes marqueurs mais non limités à ces derniers) peut être pertinent ; en effet, la perte d'information est limitée car plus de $\sim 80\%$ du génome bactérien est codant. Une raison concrète qui nous a poussé vers l'alignement sur génomes complets est la récente publication par *Almeida et al.* d'UHGG [5] (*cf.* section 2.2.3), un catalogue extrêmement complet de génomes de référence du microbiote intestinal humain, constituant une ressource précieuse et très pertinente pour notre projet.

2.1.2 . État de l'art

La SSM a récemment été suggérée comme une alternative [69] intéressante à la métagénomique *shotgun* classique, ou profond. Alors que plusieurs dizaines de millions de *reads* par échantillon sont généralement utilisés pour caractériser les échantillons de microbiote intestinal humain avec le séquençage profond [140, 143], la SSM traite généralement moins d'un million de *reads* par échantillon, ce qui réduit considérablement les coûts de séquençage, d'un facteur au moins 10. Des travaux antérieurs suggèrent que les profils taxonomiques au niveau des espèces obtenus en alignant les *reads* sur des génomes de référence étaient très similaires à ceux

obtenus avec la métagénomique *shotgun* profonde, et que les mesures d' α-diversité résultantes étaient faiblement impactées par la réduction de la profondeur de séquençage à $\sim 500K - 1M$ *reads*/échantillon [69, 150, 64, 27]. De plus, l'alignement sur des génomes de référence était plus efficace que sur un catalogue de gènes marqueurs dans le contexte de la SSM [150]. En revanche, la profondeur de séquençage requise pour l'analyse fonctionnelle dépend du niveau de granularité de l'analyse : 500K *reads*/échantillon peuvent être suffisants pour identifier les groupes d'orthologie KEGG [69], mais 3M à 5M *reads*/échantillon sont nécessaires pour détecter avec précision les gènes et les voies métaboliques [150, 175], et un séquençage très profond allant jusqu'à $\sim 60 - 80M$ *reads*/échantillon est nécessaire pour étudier les gènes de résistance aux antimicrobiens [64, 195]. Cependant, des investigations complémentaires sur la fiabilité des profils taxonomiques construits par l'alignement des *reads* SSM sur un catalogue de génomes représentatifs sont nécessaires. En effet, comme nous aurons potentiellement des profondeurs de séquençage extrêmement faibles, il est essentiel de récupérer autant d'information que possible de chaque *read*, y compris pour les *reads* ambigus qui sont alignés sur plusieurs génomes, afin d'identifier les espèces présentes en faible abondance. À notre connaissance, les travaux précédents se limitaient à l'identification des espèces cultivées, et n'incluaient pas dans les catalogues de référence des espèces non-cultivées issues de reconstructions à partir de données métagénomiques, commme les MAGs (Metagenome-Assembled Genomes). Par ailleurs, les connaissances concernant l'influence de la profondeur de séquençage sur les analyses subséquentes dans des jeux de données cliniques, comme les différences inter-groupes ou la stratification de patients, sont encore limitées et méritent d'être approfondies.

2.1.3 . Objectifs

Les objectifs de ce projet étaient de (1) construire un pipeline bioinformatique optimisé pour l'analyse de la composition du microbiote intestinal humain et (2) d'évaluer la pertinence de la SSM pour étudier le microbiote intestinal humain dans un contexte clinique.

Le pipeline bioinformatique doit permettre de construire un profil taxonomique le plus exhaustif possible, identifiant notamment tant les espèces cultivées que non-cultivées. Dans le contexte d'une faible profondeur de séquençage, il sera important d'utiliser au mieux les reads qui s'alignent sur plusieurs génomes. Enfin, la méthode ayant vocation à être utilisée dans un contexte clinique, il faut s'assurer de la qualité des profils, en maîtrisant notamment le taux de faux positifs (concernant la détection des micro-organismes) dans les profils.

Évaluer la pertinence de la SSM revient donc à évaluer l'impact de la profondeur de séquençage sur les résultats obtenus, et ainsi évaluer la perte d'information liée au passage de la métagénomique *shotgun* profonde à la SSM. Il est particulièrement intéressant d'étudier l'impact de la profondeur de séquençage au niveau des profils taxonomiques, pour évaluer le nombre d'espèces que l'on peut prétendre identifier par SSM, ainsi qu'au niveau d'une cohorte clinique, pour voir si les différences entre les patients ou les groupes de patients observées en métagénomique *shotgun* profonde sont retrouvées en SSM. Nous espérons par ce travail fournir des éléments de réponse aux scientifiques qui se demandent si cette approche peut convenir à leurs projets de recherche futurs. Ce travail est l'objet du *preprint* "*Characterizing the limits of shallow shotgun metagenomics for taxonomic profiling of human gut microbiota in clinical studies*" (https://doi.org/10.21203/rs.3.rs-1306026/v1) qui est, au moment de la rédaction de ce manuscrit, toujours en relecture au journal *Microbiome*. Pour ce chapitre, les figures étant issues pour la plupart de l'article, les légendes sont en anglais et n'ont pas été traduites.

Nous allons, dans un premier temps, travailler à partir d'un jeu de données simulé pour dé-

velopper et calibrer le pipeline bioinformatique, et évaluer formellement les profils taxonomiques obtenus, en les comparant avec les profils attendus. Ensuite, nous appliquerons notre méthode sur des jeux de données réels issus d'études précédemment publiées ayant recourt à la métagénomique *shotgun* classique (profonde) ; nous sous-échantillonnerons les *reads* pour reproduire les conditions de la SSM et étudier l'impact de la réduction de la profondeur de séquençage sur les résultats. Ce chapitre est organisé avec une première partie qui détaille le matériel et méthodes utilisées, puis une partie qui présente les résultats obtenus sur les différents jeux de données avant de s'achever sur une discussion.

2.2 . Matériel et méthodes

Dans cette section, je vais commencer par décrire les données que j'ai utilisées, d'abord les jeux de données simulés (section 2.2.2), puis les données réelles qui m'ont permis d'évaluer la méthode (section 2.2.2). Je décrirai ensuite le pipeline d'analyse que j'ai développé (section 2.2.3).

2.2.1 . Jeux de données simulés

Pour constituer des jeux de données simulés, j'ai récupéré la composition de 100 échantillons de microbiote intestinal, via la package R `curatedMetagenomicData` [133]. Ces échantillons proviennent d'une étude menée en Chine sur les maladies hépatiques [141]. Ces échantillons sont composés de 98 $\pm$ 15 espèces distinctes dans chaque échantillon et l'abondance relative s'étale entre 5.10^{-1} et 10^{-6} (la moyenne géométrique des abondances relatives des espèces vaut 5.10^{-4}). La composition de ces échantillons est donnée selon la taxonomie du NCBI. Pour pouvoir évaluer formellement les méthodes de reconstruction de profil taxonomique, il faut simuler les données à partir des génomes d'UHGG, et donc transposer la composition des 100 échantillons à la taxonomie de UHGG. Pour réaliser cela, j'ai choisi l'espèce d'UHGG la plus proche de chaque espèce des profils initiaux. Si cette transposition n'est pas parfaite, car toutes les espèces du jeu original ne sont pas présentes dans UHGG, les compositions résultantes ont la même complexité que les profils initiaux et approximativement la même composition phylogénétique.

Pour chaque espèce, j'ai choisi aléatoirement un génome parmi ceux appartenant à l'espèce. Comme seul le génome représentatif de l'espèce est présent dans le catalogue utilisé pour l'alignement, certains *reads* sont simulés à partir de génomes qui ne sont pas présents dans le catalogue. Cela reproduit la situation réaliste dans laquelle une souche que l'on retrouve dans un échantillon est différente de la souche qui représente l'espèce dans le catalogue.

Pour chaque échantillon, j'ai simulé 10M de paires de *reads* par échantillon avec *Grinder* (v0.5.3) [9], chaque *read* ayant une longueur de 125bp et les deux *reads* d'une paire étant séparé par un insert de longueur aléatoire (échantillonnée dans une distribution normale de moyenne 500bp et d'écart type 50bp), sans erreur de séquençage. J'ai ensuite sous échantillonné le jeu de données à des profondeurs de $5\ M$, $1\ M$, $500\ K$, $100\ K$, $50\ K$ and $10\ K$ *reads* par échantillon.

2.2.2 . Jeux de données réels

J'ai également utilisé trois jeux de données réels pour évaluer le pipeline d'analyse et étudier l'impact de la profondeur de séquençage sur les résultats. Les jeux de données couvrent différents contextes cliniques, différents continents et représentent des cas où les différence de composition des microbiotes entre les deux groupes d'intérêt de patients sont plus ou moins marquées ; ils

sont décrits brièvement ci-dessous. Ces jeux de données ont été produits en séquençage *shotgun* profond, et j'ai réalisé un sous-échantillonnage dans les *reads* pour reproduire les conditions de faible profondeur de séquençage.

Loomba-2017 Le jeux de données produit par Loomba *et al.* [95] contient $N = 86$ patients originaires des États-Unis souffrant de maladies hépatiques. L'objectif principal de cette cohorte est de déterminer les différences entre ceux qui ont une atteinte modérée (*N=72* patients, atteints de NALFD pour *non-alcoholic fatty liver disease*) de ceux qui ont une atteinte sévère ($N = 14$ patients, atteints de fibrose hépatique).

Matson-2018 Le jeu de données produit par Matson *et al.* [104] compare, parmi $N = 39$ patients américains atteints de mélanomes métastasés qui s'apprêtent à suivre une immunothérapie (anti-PD-1), ceux qui répondent au traitement ($N = 15$) et ceux qui n'y répondent pas ($N = 24$). Cette étude a montré des résultats très probants, notamment que chez un modèle murin, la transplantation fécale provenant des différents patients s'accompagnait d'un transfert de phénotype sur la réponse au traitement.

Qin-2014 Le jeu de données produit par Qin *et al.* [141] compare, sur des individus chinois, des patients atteints de diverses maladies hépatiques ($N = 169$) à un groupe de contrôles sains ($N = 145$). Ce jeu de données présente l'avantage d'être décomposé en une cohorte de "découverte" et une cohorte de "validation".

Sous-échantillonnage Pour reproduire les conditions de la SSM et évaluer l'effet de la profondeur de séquençage sur les résultats, j'ai réalisé un sous-échantillonnage des *reads* à partir des données de séquençage (au format `fastq`) en utilisant `seqtk`, à des profondeur de 50K,100K, 500K, 1M, 5M et 10M *reads* par échantillon. Comme mon pipeline inclut une étape de calcul de la couverture des génomes, il n'était pas possible de sous-échantillonner les *reads* directement dans la table de comptage, ce qui a rendu les opérations de sous-échantillonnage particulièrement lourdes en terme de temps de calcul.

2.2.3 . Pipeline d'analyse

Pour ce projet, j'ai mis au point un pipeline d'analyse bioinformatique dédié à la construction de profils taxonomiques à partir de données de métagénomique *shotgun* peu profonde, optimisé pour les profondeurs allant de 50K *reads* à 10M *reads* par échantillon. Le fonctionnement global de ce pipeline est présenté dans la figure 2.1 et sera détaillé dans les paragraphes suivants.

Prétraitement Les *reads* suivent d'abord un contrôle qualité qui vise à éliminer les adaptateurs de séquençage, les *reads* trop courts et de mauvaise qualité [1] (*trimmomatic* [17]). Les *reads* sont ensuite alignés sur le génome humain (`hg38`) pour retirer d'éventuels *reads* humains (contaminants).

Catalogue de référence Le choix du catalogue qui est utilisé pour réaliser l'alignement est crucial pour l'analyse, comme mentionné dans la section 2.1.1. Nous avons choisi d'utiliser le catalogue *Unified Human Gastrointestinal Genome* (UHGG, v1.0) [5]. Ce catalogue, présenté

1. `trimmomatic TRAILING:20 AVGQUAL:30 MINLEN:80`

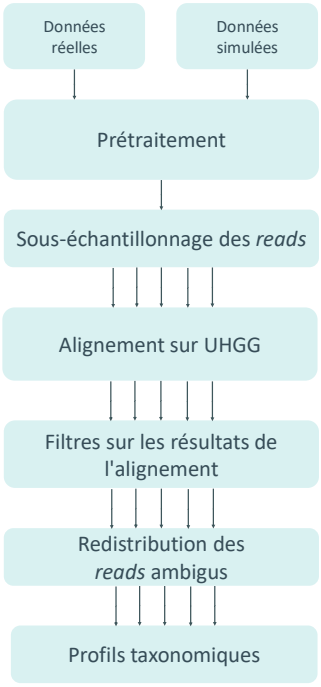


FIGURE 2.1 – Pipeline d'analyse pour l'analyse de données de métagénomique *shotgun* à faible profondeur.

dans la figure 2.2, résulte d'un travail d'agglomération de génomes provenant de différentes sources : des banques de données publiques, d'études de culturomiques (les bactéries sont isolées, cultivées puis leur génome est séquencé et assemblé) et d'études de métagénomique (les génomes sont assemblés directement à partir des *reads* métagénomiques (MAGs), *cf.* section 2.1.1). Ils ont ainsi pu répertorier plus de 200 000 génomes issus du microbiote intestinal humain. Ils ont réalisé un *clustering* de ces génomes pour les regrouper par espèce : ceux partageant au moins 95% d'identité sur au moins 30% de la longueur des génomes étaient regroupés au sein d'une même espèce. Ils ont ainsi obtenus 4 644 espèces distinctes, dont plus de 70% n'ont pas de représentant cultivé. Pour chaque espèce, ils ont désigné un génome représentatif, prenant celui qui avait la meilleure qualité (en terme de complétude et de contamination). Le catalogue contenant les génomes représentatifs des 4 644 espèces sera par la suite dénommé UHGG, et c'est sur celui-ci que j'ai aligné les *reads* de séquençage.

Le travail de Almeida *et al.* révèle l'ampleur de la partie immergée de l'iceberg que constituent les espèces non cultivées du microbiote intestinal humain. Utiliser un catalogue comme celui-ci permet d'explorer la diversité cultivée et non cultivée du microbiote, et de ne pas se restreindre aux seules espèces cultivées présentes dans les bases de données de génomes telles que RefSeq, permettant ainsi de tirer pleinement parti des données de métagénomique. Par ailleurs, par son aspect extrêmement exhaustif, il permet d'aligner entre 80% et 90% des *reads* selon les auteurs, ce qui est plus du double des résultats obtenus en utilisant RefSeq.

Outils d'alignement Pour aligner les *reads* sur le catalogue de référence, nous avons testé plusieurs outils. Nous avons comparé les résultats obtenus par `bwa mem` (`-h 50` pour récupérer jusqu'à 50 *hits*, autres paramètres laissés à leur valeur par défaut), `bwa aln` (paramètres par

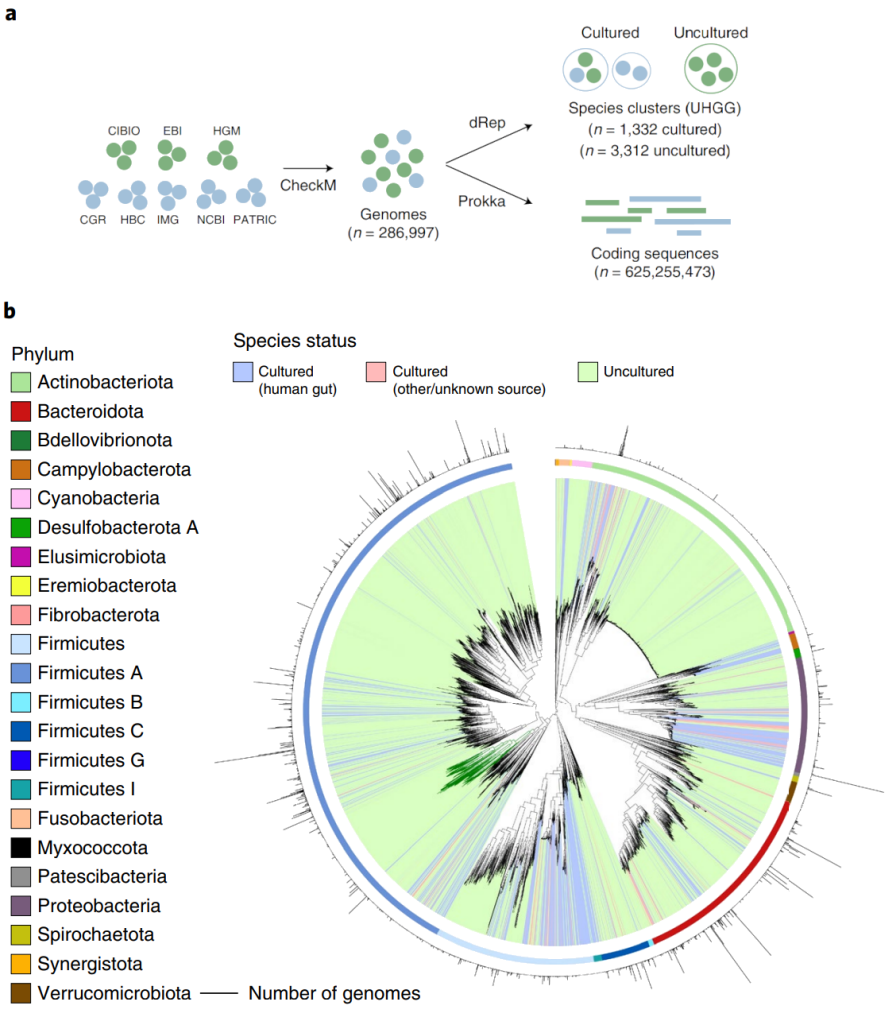


FIGURE 2.2 – Figure issue de Almeida *et al.* 2018 [5] représentant la méthodologie de construction des catalogues UHGG (pour *Unified Human Gastrointestinal Genome*) *Unified* et UHGP (pour *Unified Human Gastrointestinal Protein*) (a) ainsi que l'arbre phylogénétique des espèces présentes dans UHGG (b), colorée selon leur statut de culture (cercle intérieur), le phylum auxquelles elles ont été assignées (cercle extérieur).

défaut) et `bowtie2` en mode global (`-end-to-end`) et local (`-local`).
Les *reads* qui sont alignés sur plusieurs génomes du catalogue sont désignés comme "ambigus", ceux alignés sur un seul génome sont désignés comme "non-ambigus" et les autres sont désignés comme "non-alignés". Les *reads* ambigus proviennent de région conservées entre les espèces, ou encore d'éléments mobiles.
Pour chaque génome représentatif d'UHGG et en utilisant `samtools`, nous avons calculé (1) le nombre de *reads* qui s'alignent sur celui-ci, noté RC pour *read counts*, ainsi que (2) la fraction du génome couverte par des *reads*, définie comme le nombre de positions du génome couvertes par au moins un *read* divisé par la taille du génome et notée FC pour *fraction covered*. De manière similaire, nous avons mesuré uRC et uFC qui correspondent aux nombres de *reads* non-ambigus et à la fraction des génomes couverte par des *reads* non-ambigus (le u correspondant à *unambiguous*). Nous avons également mesuré, pour chaque génome, le ratio entre le nombre de *reads* ambigus et non-ambigus, appelé ratio de spécificité ($SR = \frac{uRC}{RC}$).

Lors des simulations, puisqu'on connaît la composition théorique des échantillons, un *read* est désigné comme vrai positif (ou *TP* en anglais) si le génome sur lequel il s'aligne (ou au moins un des génomes sur lequel il s'aligne dans le cas où le *read* est ambigu) fait partie des génomes attendus.

Estimation des abondances relatives des espèces Pour chaque espèce s, nous avons calculé la couverture moyenne de son génome représentatif $C_s = \frac{1}{\ell_s} \sum_i r_{i,s}$, avec ℓ_s la longueur du génome représentatif de l'espèce s et $r_{i,s}$ la longueur du read i qui a été aligné de manière non ambiguë sur le génome représentatif de s. Une première estimation de l'abondance relative des espèces est obtenue en divisant la couverture moyenne de chaque génome représentatif par la somme des couvertures de tous les génomes représentatifs : $A_s = \frac{C_s}{\sum_j C_j}$.

Redistribution des *reads* ambigus Nous avons ensuite affiné l'estimation des abondances relatives des espèces en redistribuant les *reads* ambigus. Chaque *read* ambigu est assigné aléatoirement à un des génomes représentatifs sur lesquels il a été aligné, avec une probabilité proportionelle aux abondances relatives des espèces, estimées à partir des seuls *reads* non-ambigus. Ainsi, nous calculons C'_s puis A'_s qui tiennent compte des *reads* ambigus redistribués. Nous avons également testé une approche dans laquelle les *reads* ambigus étaient partagés entre les génomes sur lesquels ils étaient alignés, avec des contributions proportionnelles aux abondances relatives des espèces A_s. Les résultats étaient tellement identiques que seule la première méthode sera montrée dans les résultats, par souci de simplicité.

Filtrage des résultats d'alignement L'alignement de *reads* sur un catalogue de génomes représentatifs tel que UHGG produit des résultats extrêmement bruités. En effet, si on prend les données brutes de l'alignement, des reads s'alignent sur de nombreux génomes absents des échantillons. Même s'ils ne sont pas des faux positifs au sens de l'alignement en lui-même, car les séquences des *reads* sont bien présentes dans les génomes de référence, ces *hits* sont de faux positifs au sens de la construction du profil taxonomique, et seront donc dénommés ainsi pour la suite. Ces alignements sont généralement concentrés sur des régions du génomes qui sont conservées entre espèces. Si de nombreux faux positifs peuvent être facilement éliminés en appliquant un seuil sur le nombre de *reads* car ils sont identifiés par peu de *reads* (RC et uRC très bas), d'autres sont plus difficiles à détecter, et il est particulièrement intéressant, dans un contexte de faible profondeur de séquençage, d'optimiser cette étape. Même s'il y a toujours un compromis à faire entre précision [2] et rappel [3] pour filtrer les résultats de l'alignement, on cherche à identifier les meilleures stratégies, l'impact de cette étape de filtrage dans les profils taxonomiques résultants. Cette étape constitue une tâche de classification, comme présenté dans la table 2.1.

Nous avons d'abord filtré les données en utilisant des seuils sur le nombre de *reads* alignés (RC et uRC), ainsi que sur la *fraction couverte* (FC et UFc), tel qu'il est classiquement réalisé. Nous avons ensuite utilisé des méthodes multivariées, qui combinent ces informations pour réaliser une tâche de classification. Nous avons utilisé trois méthodes de classification différentes : une régression logistique, une analyse discriminante linéaire (en anglais, *linear discriminant analysis* ou LDA) et enfin des forêts aléatoires. La régression logistique et la LDA sont des modèles

2. $\frac{TP}{TP+FP}$
3. $\frac{TP}{TP+FN}$

Génome Id	RC	uRC	FC	uFC	SR	Catégorie
GUT_GENOME226406	2090	1295	0.020	0.017	0.620	FP
GUT_GENOME283627	1167	120	0.023	0.009	0.103	FP
GUT_GENOME096174	17084	2920	0.148	0.060	0.171	TP
GUT_GENOME096080	11005	1117	0.088	0.040	0.101	TP
...			...			

TABLE 2.1 – Exemple des données récupérées à partir de l'alignement. Pour chaque génome représentatif de UHGG, on calcule les comptages obtenus avec tous les *reads* (RC) ou seulement les *reads* non-ambigus (uRC) ainsi que les fractions couvertes par tous les *reads* (FC) ou uniquement par des *reads* non-ambigus (uFC) et enfin le ratio de spécificité. La dernière colonne du tableau décrit si l'espèce est présente dans l'échantillon (TP pour *true positive*) ou non (FP pour *false positive*), une information connue dans les simulations.

linéaires qui peuvent s'interpréter comme la projection des données sur des axes qui permettent de discriminer au mieux les données. La régression logistique détermine cet axe en maximisant la vraisemblance des données alors que la LDA cherche à maximiser le rapport entre la variabilité inter-groupes et la variabilité intra-groupes [136]. La forêt aléatoire est un algorithme d'apprentissage automatique qui construit un ensemble d'arbres de décision, chacun utilisant une sous-partie des données pour déterminer la classe d'appartenance. Pour classer de nouvelles données, la probabilité d'appartenance aux classes est déterminée en prenant la moyenne de la prédiction de l'ensemble des arbres (*i.e.* de la forêt) [20, 22].

Les filtres ont été mis au point et évalués sur les jeux de données simulés présentés dans la section 2.2.1. Concernant la régression logistique, la LDA et la forêt aléatoire, comme ces méthodes reposent sur un ensemble d'apprentissage, j'ai utilisé une validation croisée, en séparant le jeux de données en 4 groupes d'échantillons de tailles égales, et successivement entraîné les classifieurs sur trois quarts des données, puis évalué sur le quart restant. Le but étant de développer un filtre applicable aux échantillons réels dont on ne maîtrise pas la profondeur de séquençage, les modèles de classification sont entraînés et évalués avec des jeux de données comprenant toutes les profondeurs de séquençage. Le nombre de *reads* total de l'échantillon est donné comme une variable au classifieur, en plus des autres.
Quelques soient les filtres utilisés, il y a un seuil à appliquer qui affecte la stringence du filtre et agit ainsi sur le compromis entre le taux de faux négatifs (élevé si le filtre est très stringent) et le taux de faux positifs (élevé si le filtre est peu stringent). Pour pouvoir comparer les filtres entre eux, j'ai utilisé comme critère l'AUC (*Area Under the Curve*) de la courbe ROC (*Receiver Operating Characteristic*) correspondant à la tâche de classification décrite précédemment qui permet d'évaluer la performance de classification en explorant toutes les valeurs du seuil. J'ai également déterminé, pour chaque méthode et à chaque profondeur de séquençage, le seuil qui permet de tolérer un taux FDR[4] à 0.1 dans les profils. À ce seuil, on maîtrise le taux de faux positifs et on peut ainsi comparer, entre les méthodes et les profondeurs de séquençage, le taux de faux négatifs pour étudier quelles méthodes permettent de récupérer le plus d'espèces effectivement présentes.
Sur les données réelles, nous avons utilisé (1) la meilleure stratégie de séquençage résultant de

4. $FDR = \frac{FP}{TP+FP}$

l'analyse précédemment décrite, avec des seuils dépendants de la profondeur de séquençage, et (2) un filtre basique, qui élimine toutes les espèces avec une abondance relative inférieure à 10^{-4}, avec une fraction couverte inférieure à 10^{-2} et une fraction couverte par des *reads* non ambigus inférieure à 10^{-4}. Ce filtre basique est similaire à celui utilisé par Santiago-Rodriguez *et al.* [150]. Il faut noter que ce filtre est relativement permissif, et que les seuils sont invariants à la profondeur de séquençage.
N.B. : dans la suite de ce chapitre, le terme *filtre basique* fera référence à ce filtre, et non pas aux filtres univariés.

2.3 . Résultats

Je présenterai dans un premier temps les résultats issus des jeux de données simulés, qui m'ont permis de mettre au point le pipeline d'analyse, avec en préambule le calibrage de l'alignement (section 2.3.1) puis l'importance des stratégies de filtrage des résultats de l'alignement (section 2.3.2), et dans un deuxième temps les résultats issus de données réelles et étudiant l'impact de la profondeur de séquençage sur les profils taxonomiques (2.3.3) et sur les analyses statistiques de jeux de données (2.3.4).

2.3.1 . Calibrage de l'étape d'alignement

Lorsqu'on aligne des données de métagénomique sur des catalogues de génomes complets, on est souvent contraint pour des raisons de temps de calcul de choisir un seul génome représentatif pour chaque espèce, comme c'est le cas d'UHGG. Par conséquent, les *reads* sont alignés sur des génomes qui sont rarement issus de la même souche, mais plutôt d'une autre souche de la même espèce. Ainsi, il est important de choisir au mieux les paramètres d'alignement, afin d'être suffisamment souple pour tenir compte de la diversité intra-espèce (plus de 95% d'identité) mais pas trop, afin de limiter le nombre de faux positifs.

J'ai dans un premier temps, sur un sous-ensemble du jeu de données simulé présenté dans la section 2.2.1, comparé différents outils d'alignements, pour évaluer lequel donnait les meilleurs résultats. J'ai distingué les cas où les *reads* proviennent du génome représentatif (figure 2.3, gauche) et les cas où ils proviennent d'un autre génome (figure 2.3, droite). Le deuxième cas est plus réaliste, mais également plus difficile en raison (1) des différences avec le génome de référence sur les régions communes et (2) des gènes spécifiques au génome de la souche qui seront absents du génome de référence. On observe sur la figure 2.3 que l'alignement est bien meilleur si le génome dont sont issus les *reads* est le génome représentatif de l'espèce. Dans le cas inverse, le génome d'où proviennent les *reads* n'est pas présent dans le catalogue, ce qui entraîne plus de *reads* non-alignés et plus de faux positifs. On note que `bowtie2`, qu'il soit utilisé en mode local ou global, aligne moins de *reads* que `bwa mem` et `bwa aln`. J'ai exploré différentes valeurs de seuil de score minimal pour reporter un alignement dans les paramètres de `bowtie2`, dans l'optique d'avoir moins de *reads* non-alignés, mais la diminution du seuil s'accompagnait d'une forte augmentation du taux d'alignement ambigu, ce qui était contre-productif pour la résolution des profils taxonomiques. Nous avons finalement choisi `bwa mem`, qui offrait le meilleur compromis entre les taux d'alignement, d'alignements ambigus et de faux positifs.

Concernant les alignement ambigus, on observe sur la figure 2.4 que la grande majorité des *reads* ont peu de *hits*. Le nombre de *reads* alignés sur plus de 3 génomes est marginal. On observe sur la figure 2.4 que lorsque les *reads* sont alignés sur plusieurs génomes, ces derniers

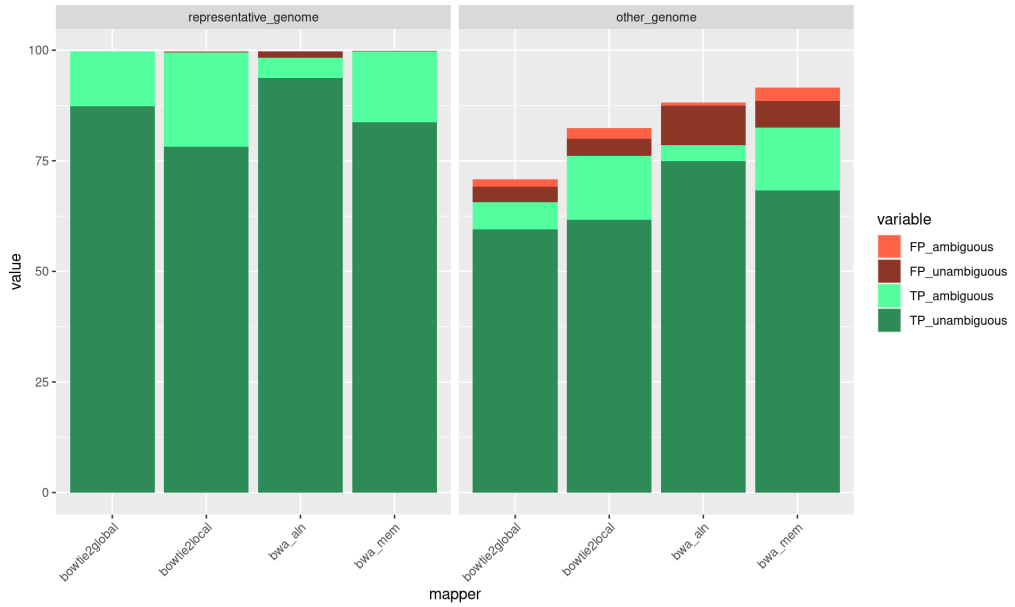


FIGURE 2.3 – Comparaison des outils d'alignement. Dans le panneau de gauche, les *reads* proviennent du génome représentatif de l'espèce présent dans le catalogue UHGG utilisé pour l'alignement, alors que dans le panneau de droite, les *reads* proviennent d'un autre génome de la même espèce, et donc absent du catalogue. Les *reads* sont désignés comme *TP_unambiguous* s'ils sont alignés uniquement sur le génome représentatif de la bonne espèce, et *FP_unambiguous* s'ils sont alignés uniquement sur un autre génome ; les *reads* étant alignés sur plusieurs génomes sont désignés comme *TP_ambiguous* si le génome attendu fait parti des génomes sur lesquels le *reads* est aligné, et *FP_ambigous* s'il n'en fait pas partie. La hauteur totale des barres représente le pourcentage de *reads* alignés.

appartiennent souvent au même genre. Cela confirme la cohérence de l'alignement et de la taxonomie du catalogue utilisé.

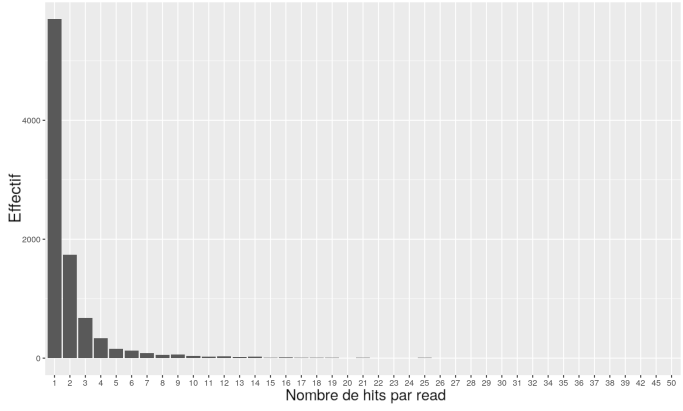


FIGURE 2.4 – Distribution du nombre de *hits* par *read* avec `bwa mem`.

Nous avons par ailleurs étudié l'effet de la redistribution des *reads* ambigus pour affiner les profils taxonomiques (cf section 2.2.3). Nous avons regardé, dans les données simulées uniquement, les vrais positifs, c'est-à-dire les populations attendues et observées dans les profils. On observe sur la figure 2.6 que la redistribution des *reads* ambigus améliore la corrélation entre

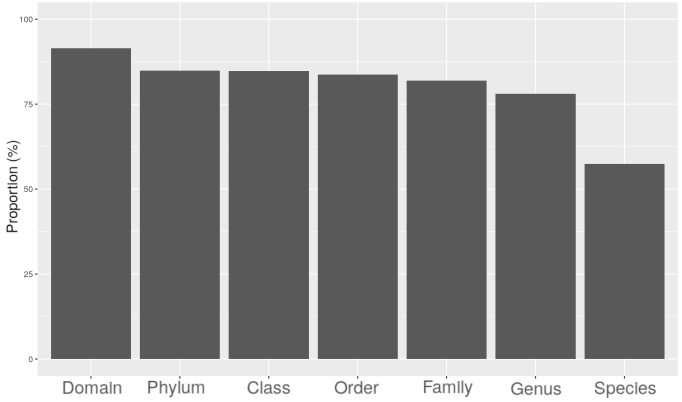


FIGURE 2.5 – Nombre de *reads* pouvant être assignés aux différents niveaux taxonomiques. Seuls les *reads* n'étant alignés que sur un seul génome sont assignés jusqu'à l'espèce, les autres sont assignés jusqu'au rang du dernier ancêtre commun des différents génomes sur lesquels ils sont alignés.

les abondances relatives attendues et estimées, et ce quelque soit la profondeur de séquençage.

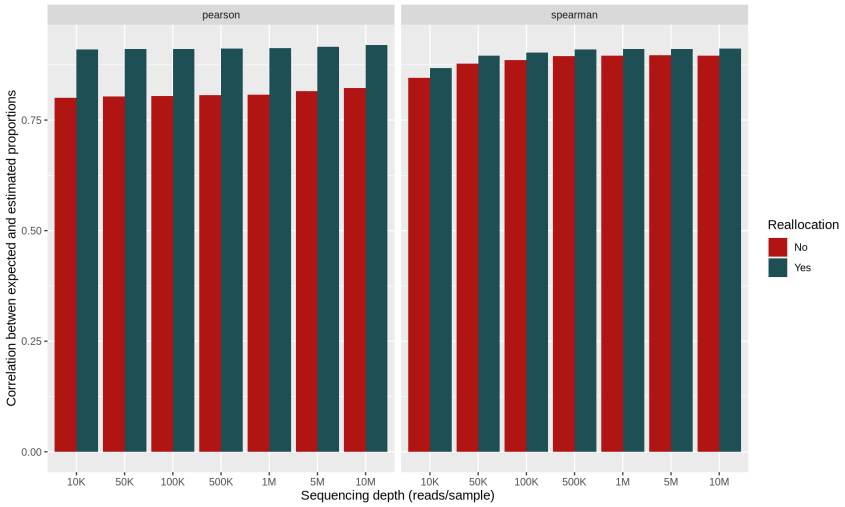


FIGURE 2.6 – Corrélations de Pearson et de Spearman entre les abondances relatives estimées et attendues des espèces sans redistribution des *reads* ambigus (en rouge) et avec redistribution (en vert), aux différentes profondeurs de séquençage. Seuls les vrais positifs (espèces attendues et observées) sont pris en compte pour établir les corrélations des abondances relatives.

2.3.2 . Stratégies de filtrage des résultats de l'alignement

Dans cette section, nous allons dans un premier temps comparer entre eux les différents filtres proposés en section 2.2.3 pour déterminer la stratégie de filtrage la plus performante. Ensuite, nous regarderons plus en détail l'impact de la profondeur de séquençage sur les profils taxonomiques résultants du meilleur filtre, pour évaluer les limites de l'approche SSM.

Comparaison des filtres Sur les données brutes issus de l'alignement, on observe de très nombreux faux positifs, jusqu'à 0.92 de FDR[5]. Nous avons cherché dans un premier temps à appliquer des seuils indépendamment sur les différents indicateurs mesurés sur chaque génome représentatif des espèces, à savoir le comptage des reads (RC) et fraction du génome couverte (FC), pour les reads totaux et non-ambigus, et présentés dans la table 2.1 (page 33).

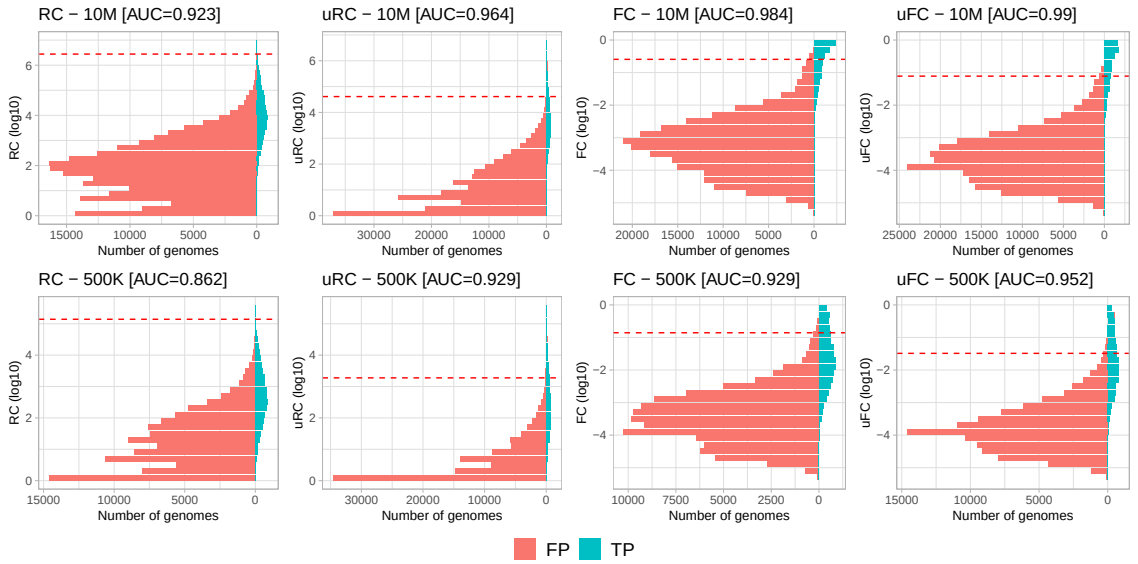


FIGURE 2.7 – Histogrammes des comptages de *reads* totaux et non-ambigus (RC et uRC) et les fractions du génome couvertes par les *reads* totaux et non-ambigus (FC et uFC) pour deux profondeurs de séquençage : 10M (haut) et 500K reads (bas). Les histogrammes relatifs aux faux positifs (FP) sont représentés en rouge et ceux relatifs aux vrais positifs (TP) en bleu. Les lignes horizontales en rouge représentent les valeurs seuils qui assurent un taux de *FDR* inférieur à 0.1 dans les données. La majorité des faux positifs ont une faible couverte et/ou une faible profondeur.

On observe sur la figure 2.7 que les quatre indicateurs sont généralement plus faibles pour les faux positifs que pour les vrais positifs, comme attendu. Cependant, les distributions se superposent, ne permettant pas de fixer un seuil qui puisse discriminer de manière satisfaisante les vrais positifs et les faux positifs. En regardant les valeurs des AUCs, on observe qu'utiliser un seuil sur les comptages de *reads* donne de moins bons résultats que d'utiliser un seuil sur les fractions couvertes, et qu'utiliser seulement les *reads* non-ambigus (*uRC* et *uFC*) fonctionne mieux que d'utiliser tous les *reads* (*RC* et *uRC*). On note également, en regardant la valeur des AUCs, que différencier les vrais positifs et les faux positifs est plus facile à 10M de *reads* par échantillon qu'à 500K *reads* par échantillon, quel que soit l'indicateur utilisé. Les résultats montrent l'importance du choix du filtre pour les données SSM.

5. $FDR = \frac{FP}{TP+FP}$

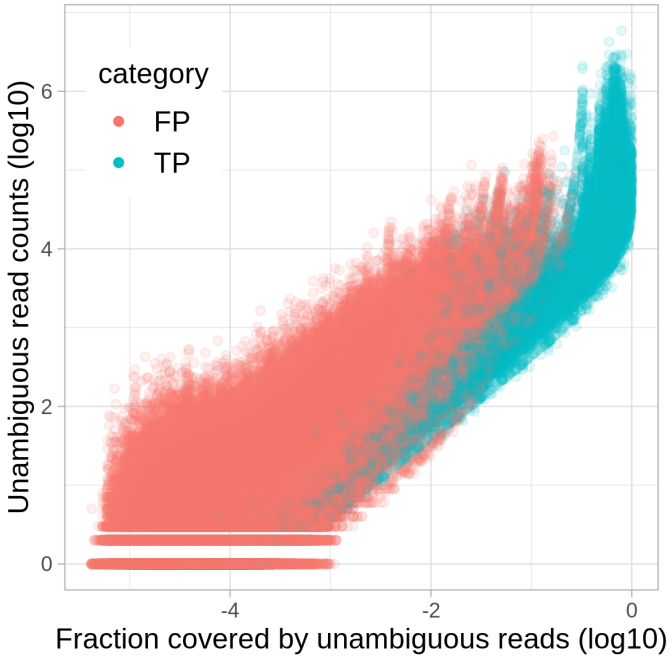


FIGURE 2.8 – Fraction couverte par les *reads* non-ambigus (uFC) et comptage de *reads* non-ambigus (uRC) du génome représentatif de chaque espèce identifiée lors de l'alignement, que celles-ci soient présentes (vrais positifs, en bleu) ou absentes (faux positifs, en rouge) des profils théoriques.

Nous avons ensuite entraîné des classifieurs pour réaliser la tâche de manière plus fine. On observe en effet dans la figure 2.8 que les nuages de points correspondants aux vrais positifs (en bleu) et aux faux positifs (en rouge) sont dissociables. On note que l'application de seuils univariés, qui correspondent dans cette représentation à des droites horizontales et/ou verticales au dessus (et/ou à droite) desquels tous les génomes sont considérés comme présents, est sous-optimale. En effet, au-delà des vrais positifs évidents, dont plus de 10% du génome est couvert, et des faux positifs évidents dont moins de 0.1% du génome est couvert, il y a de très nombreux génomes identifiables avec une stratégie appropriée. Dans cette zone intermédiaire, les vrais positifs correspondent aux génomes dont la couverture est élevée malgré un nombre de *reads* relativement faible. Ainsi, les vrais positifs ont des *reads* qui sont plutôt bien répartis le long du génome, alors que les faux positifs, même s'ils ont beaucoup de *reads*, ne sont couverts que dans certaines zones. Notre hypothèse est que les méthodes d'apprentissage automatique pourront permettre de tirer parti de ces observations, en s'appuyant sur les deux critères simultanément plutôt que des les considérer séparément.

Les classifieurs que nous avons utilisés (*cf.* section 2.2.3) prennent en entrées le nombre de *reads* non-ambigus alignés sur les génomes, les fractions couvertes par des *reads* non-ambigus, le ratio de spécificité ($SR = \frac{uRC}{RC}$) et la profondeur de séquençage (nombre total de *reads* de l'échantillon). Le ratio de spécificité s'est avéré ne pas être un bon indicateur s'il était considéré individuellement mais améliorait la classification lorsqu'on le rajoutait aux autres. Comme les classifieurs prennent la profondeur de séquençage en entrée, ils sont ainsi utilisables à toute les profondeurs de séquençage. Cependant, nous avons utilisé des seuils distincts pour chaque profondeur de séquençage, établis sur le jeu de données de test, pour assurer un taux de *FDR* inférieur à 0.1.

Méthode	AUC		FNR	
	Entraînement	Test	Entraînement	Test
LDA	0.947 ± 0.001	0.947 ± 0.002	0.415 ± 0.002	0.416 ± 0.010
Régression logistique	0.958 ± 0.001	0.958 ± 0.002	0.388 ± 0.003	0.389 ± 0.013
Forêt aléatoire	0.999 ± 0.0001	0.969 ± 0.003	0.037 ± 0.001	0.292 ± 0.008

TABLE 2.2 – Performance des classifieurs en termes d'AUC et *FNR* lorsqu'on calibre la méthode pour avoir un *FDR* inférieur à 0.1. Ces indicateurs de performance sont mesurés sur les jeux d'entraînement et de test, lors d'une validation croisée qui utilise successivement 3/4 des données pour l'entraînement et 1/4 pour le test. Les valeurs données dans le tableau représentent les moyennes $\pm$ écarts-types à travers les 4 répétitions.

Les modèles de LDA et régression logistique donnent une AUC globale de 0.947 et 0.958 respectivement (*cf.* table 2.2), avec une excellente capacité de généralisation car les performances sont identiques sur le jeu d'entraînement et sur le jeu de validation. Les résultats donnés par les forêts aléatoires sont encore meilleurs, avec une AUC proche de 1 sur les jeux d'entraînement et de 0.969 sur les jeux de test. Les forêts aléatoires ont donc tendance à sur-ajuster la classification aux données, mais la performance de classification sur le jeu de test restant supérieure aux autres méthodes, c'est celle-ci qui a été retenue pour la suite.

On observe sur la figure 2.9 que toutes les stratégies de filtre qui utilisent les classifieurs, et donc plusieurs métriques simultanément, améliorent la qualité des profils, et ce à toutes les profondeurs de séquençage. Globalement, les AUCs sont très élevées et les différences entre elles sont faibles, surtout aux profondeurs de séquençage les plus élevées, car il y a une grande masse de faux positifs évidents qui sont correctement identifiés par tous les filtres. En revanche, lorsque l'on regarde la différence de *FNR* aux seuils qui assurent un taux de *FDR* inférieur à 0.1, on observe nettement l'intérêt des filtres utilisant les classifieurs et surtout des forêts aléatoires, qui parviennent à identifier beaucoup plus d'espèces.

Comparaison avec le filtre basique A titre de comparaison, le filtre basique, qui retire les espèces les moins abondantes et les moins couvertes (*cf.* fin de la section 2.2.3 pour les détails) donne des résultats qui sont très loin d'être satisfaisants, avec un FDR de 0.44 (près d'une espèce identifiée sur deux est un faux positif) et un FNR de 0.46 (près d'une espèce à identifier sur deux est absente du profil), sur l'ensemble du jeu de données simulé (données agrégées sur les 100 échantillons à toutes les profondeurs de séquençage envisagées). De manière très intéressante, parmi les espèces qui sont retenues par le filtre basique et qui ne sont pas retenues par le filtre basé sur la forêt aléatoire, 89% sont des faux positifs. À l'inverse, parmi les espèces retenues par le filtre basé sur la forêt aléatoire et qui ne sont pas retenues par le filtre basique, seuls 17% sont de faux positifs. Ces résultats, ainsi que ceux présentés dans la table 2.2, montrent que les profils résultants du filtre basé sur les forêts aléatoires sont plus proches de la réalité que ceux résultant du filtre basique.

Impact de la profondeur de séquençage sur les faux négatifs En utilisant dans cette section le filtre par forêt aléatoire avec des seuils assurant à chaque profondeur de séquençage un taux de *FDR* inférieur à 0.1, j'ai étudié les profils taxonomiques résultants en terme de faux négatifs, c'est-à-dire les populations attendues mais absentes des profils. On observe sur la figure 2.10 que, malgré une stratégie de filtre optimisée, plus on réduit la profondeur de séquençage, plus

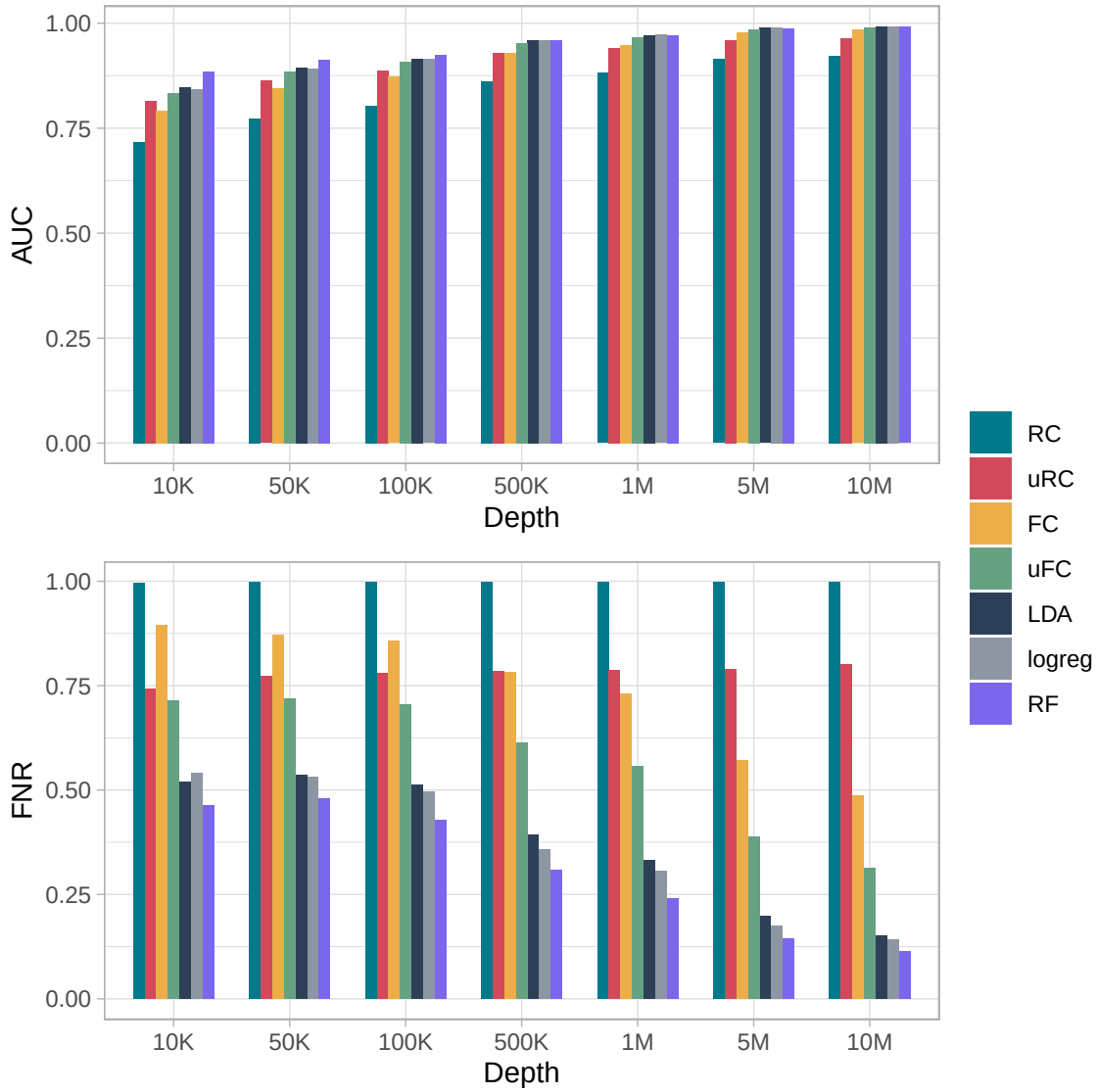


FIGURE 2.9 – Performances des différentes stratégies de filtrage en fonction de la profondeur de séquençage : AUC (en haut) et taux de faux négatifs (FNR, en bas) au seuil qui assure un taux de *FDR* inférieur à 0.1. Les stratégies de filtre reposent soit sur un seuil sur chaque indicateur pris indépendamment (RC, uRC, FC et uFC) soit sur des classifieurs (LDA, régression logistique notée *logreg* et forêt aléatoire notée *RF*). Concernant les méthodes filtres utilisant un classifieur, les valeurs représentées sont les moyennes sur les 4 itérations de la validation croisée.

les faux négatifs sont nombreux. On note que les différences entre les faux négatifs à 10M *reads* par échantillon et ceux à 1M ou 500K concernent principalement les espèces rares, d'une abondance relative inférieure à 10^{-3}. Pour donner des repères numériques, à une profondeur de 500K *reads* par échantillon, 90% des espèces présentes à une abondance relative au dessus de 4.10^{-4} sont détectées, alors que cet indicateur est à 2.10^{-4} à une profondeur de 1M *reads* par échantillon, et 3.10^{-5} à 5M *reads* par échantillon.

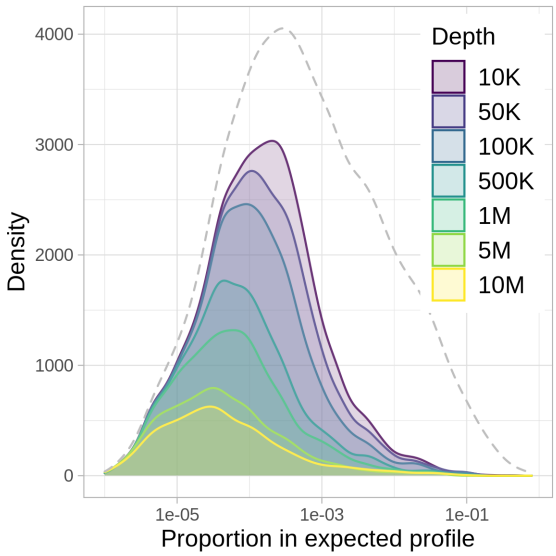


FIGURE 2.10 – Répartition des faux négatifs en fonction de leur abondance relative théorique, dans les profils résultants du filtre par forêt aléatoire assurant un taux de FDR inférieur à 0.1. La courbe en pointillé correspond à l'ensemble total des espèces à identifier. Les faux positifs sont concentrés dans les espèces faiblement abondantes.

Après avoir utilisé les simulations pour développer et évaluer le pipeline d'analyse dédié à la construction de profils taxonomiques à partir de données de SSM, les deux sections qui suivent ont pour but d'évaluer, en utilisant les jeux de données réelles présentés dans la section 2.3.4, l'impact de la profondeur de séquençage sur les profils taxonomiques et la stratification de patients, pour évaluer les forces et faiblesses de la SSM.

2.3.3 . Impact de la profondeur de séquençage sur les profils taxonomiques

L'application du filtre par forêt aléatoire permet d'obtenir des profils d'une richesse moyenne de 128 ± 26 espèces par échantillon à pleine profondeur (moyenne $\pm$ écart-type). La richesse en espèces diminue graduellement et ne vaut plus que 45 ± 21 espèces/échantillon à une profondeur de 500K *reads* par échantillon (*cf.* figure 2.11A). Si la richesse en espèces diminue avec la profondeur de séquençage, on observe sur cette même figure, la linéarité entre la richesse observée à pleine profondeur et celle observée à très faible profondeur, même jusqu'à 50K *reads* par échantillon. Contrairement à la richesse en espèces, l'indice de *Shannon*, qui tient compte des abondances relatives des espèces et donne plus de poids aux espèces majoritaires, est moins impacté par la profondeur de séquençage (*cf.* figure 2.11B), ce qui témoigne du fait que les espèces perdues à faible profondeur sont les plus rares, comme vu dans la figure 2.10. En étudiant la distance de Bray-Curtis entre un échantillon à faible profondeur et sa référence, définie comme le même échantillon à pleine profondeur, on voit sur la figure 2.11C que la distance augmente graduellement lorsqu'on diminue la profondeur de séquençage, ce qui confirme l'influence de la profondeur de séquençage sur les profils taxonomiques.

Comparativement, le filtre basique est beaucoup plus permissif, produisant des profils taxonomiques avec une plus grande richesse en espèces et moins impactés par la profondeur de séquençage (*cf.* figure 2.11D et E). Les distances par rapport aux échantillons à pleine profondeur sont plus faibles, surtout jusqu'à 500K *reads*/échantillon (*cf.* figure 2.11F).

Ces résultats montrent le fort impact de la profondeur de séquençage sur les profils taxonomiques. Plus la profondeur de séquençage est faible, moins on peut identifier d'espèces. L'impact de la profondeur de séquençage est plus important en utilisant le filtre basé sur des forêts aléatoires, qui assure la fiabilité des profils, qu'avec le filtre basique. Pour rappel, on avait mis en évidence dans la section 2.3.2, que le filtre basique produisait des profils de mauvaise qualité (taux de faux positifs et faux négatifs élevés). On observe ici qu'il produit des profils qui sont plus riches et moins impactés par la profondeur de séquençage que le filtre basé sur la forêt aléatoire.

2.3.4 . Impact de la profondeur de séquençage sur la stratification des patients

Dans cette section, nous allons nous intéresser à la capacité de la métagénomique *shotgun* à faible profondeur à retrouver les différences entre les groupes de patients, à l'échelle d'un jeu de données complet. Pour chacun des 3 jeux de données considérés (*cf.* section 2.2.2), qui présentent des niveaux de signal très variables, nous avons comparé les résultats obtenus à pleine profondeur et ceux obtenus à faible profondeur.

Pour pouvoir comparer équitablement les résultats statistiques obtenus aux différentes profondeurs de séquençage, il était nécessaire de restreindre l'analyse aux échantillons pour lesquels nous avons des données à toutes les profondeurs de séquençage comparées. Ainsi, dans Loomba-2017, seuls les 77 échantillons (sur 86) qui possèdent plus de 10M *reads* par échantillon ont été utilisés pour les analyses qui vont suivre, dans Matson-2018 les 39 échantillons ont été utilisés, et dans Qin-2014 nous avons utilisé les 266 échantillons qui possèdent plus de 5M *reads* par échantillon sur les 314 échantillons du jeu de données. Pour ce dernier jeu de données, l'analyse n'a pas été menée à 10M *reads* par échantillon car cela aurait nécessité d'exclure trop d'échantillons (142 échantillons sur les 314 de départ).

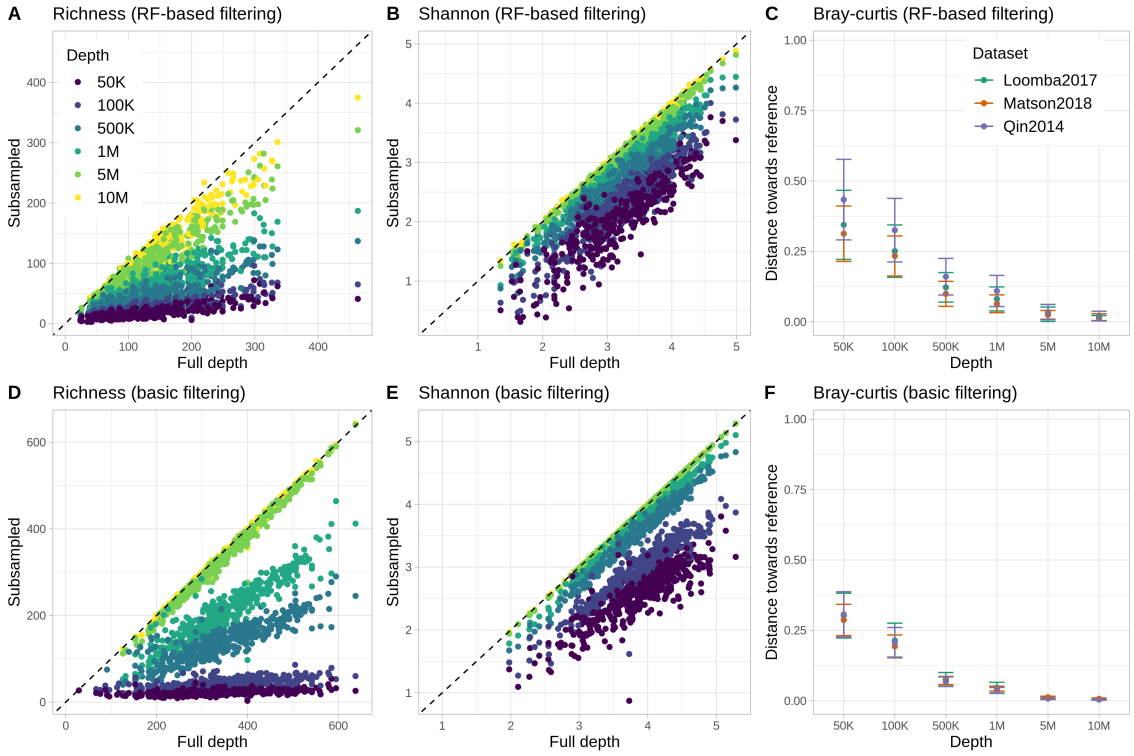


FIGURE 2.11 – Impact de la profondeur de séquençage sur les profils taxonomiques dans les échantillons provenant des 3 jeux de données : richesse observée, indice de Shannon et distance de Bray-Curtis entre les données à pleine profondeur et celles sous-échantillonnées, aux différentes profondeurs de séquençage, en utilisant le filtre basé sur les forêts aléatoires (A, B et C respectivement) et le filtre basique (D, E et F respectivement). Les valeurs sur les figures C et F représentent la distance moyenne sur l'ensemble des échantillons de chaque jeu de données et les barres d'erreur représentent l'écart-type de ces distances.

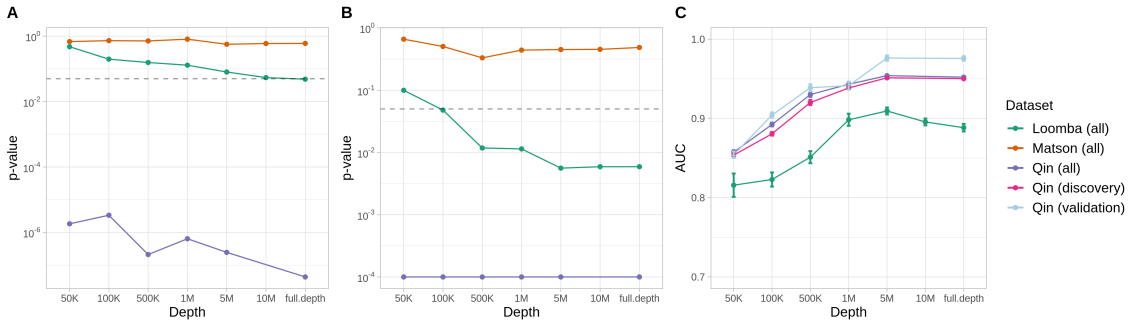


FIGURE 2.12 – Impact de la profondeur de séquençage sur les différences entre les groupes de patients dans les différents jeux de données : significativité de la différence inter-groupes en termes (A) d'α-diversité (indice de Shannon, test de Wilcoxon) et (B) de β-diversité (PERMANOVA sur la matrice de distances de Bray-Curtis). Performance de classification (C) des patients selon leur groupe (classification par forêts aléatoires basée sur la composition de leur microbiote, *cf.* paragraphe 2.3.4 pour les détails). Les barres d'erreurs représentent l'écart-type de l'AUC à travers 10 répétitions du procédé de classification.

Différences d'α-diversité Nous avons testé la significativité de la différence entre les groupes de patients obtenus par les différentes études par un test de Wilcoxon sur les indices de Shannon. Comme attendu à la vue des résultats au niveau des profils taxonomiques (figures 2.11A et 2.11D), les différences entre groupes en termes d'α-diversité sont peu impactées par la profondeur de séquençage (voir figure 2.12A). Une différence très significative entre groupes est observée dans Qin-2014 quelque soit la profondeur de séquençage. Dans Matson-2018, il n'y a pas de différence entre les groupes et la réduction de profondeur n'affecte pas ce résultat. Dans Loomba-2017, il existe une tendance légère mais non significative, y compris à pleine profondeur, et on note une dégradation du signal en dessous de 5M *reads* par échantillon.

Différences de β-diversité La différence entre les composition des microbiotes des différents groupes a été évaluée par un test de PERMANOVA, qui évalue l'effet d'une co-variable (ici le groupe d'appartenance des échantillons) sur la matrice de distances entre les échantillons. On observe sur la figure 2.12B que la significativité de ce test statistique est peu impactée par la profondeur de séquençage. Là encore, le signal est très fort dans le jeu de données de Qin-2014, une différence plus modérée mais très bien conservée jusqu'à 500K *reads*/échantillon dans les données de Loomba-2017, et aucun signal dans Matson-2018. On peut supposer que certaines populations clés de la différence inter-groupes dans Loomba-2017 sont perdues en dessous de 500K *reads* par échantillon.

Recherche de biomarqueurs La recherche de biomarqueurs, c'est-à-dire d'espèces différentiellement abondantes dans les deux groupes, a été menée par un test de Wilcoxon sur les abondances relatives des espèces, avec une correction de la *p-value* par la méthode de Benjamini-Hochberg. Je n'ai pas montré les résultats de cette analyse sur Matson-2018 car aucune espèce ne ressortait. Concernant Loomba-2017, comme le signal était faible, j'ai utilisé un seuil permissif ($FDR < 0.1$) pour reporter suffisamment d'espèces et ainsi pouvoir évaluer l'effet de la profondeur de séquençage. Concernant Qin-2014, comme le signal était très fort et qu'il y avait une cohorte de découverte et une de validation, je n'ai reporté que les espèces qui avaient un $FDR < 0.05$ dans les deux cohortes. On observe sur la figure 2.13 que le nombre de biomarqueurs identifiés décroît avec la profondeur de séquençage. Par exemple, parmi les 6 biomarqueurs identifiés dans Loomba-2017 à pleine profondeur, seuls les 4 qui étaient retrouvés à une abondance relative moyenne[6] supérieure à 10^{-2} sont retrouvés à 500K *reads* par échantillon. De manière similaire, dans Qin-2014, parmi les 56 biomarqueurs identifiés à pleine profondeur, les 2 qui ont une abondance relative moyenne supérieure à 10^{-2}, et 12 des 37 qui ont une abondance relative moyenne entre 10^{-2} et 10^{-3} sont retrouvés à 500K *reads* par échantillon. Dans les deux jeux de données, aucun des biomarqueurs les plus rares, c'est à dire dont l'abondance relative moyenne est inférieure à 10^{-3}, n'est retrouvé à cette profondeur.

Classification des patients Enfin nous avons réalisé une classification des patients selon leur groupe d'intérêt, puis mesuré l'impact de la profondeur de séquençage sur la performance de classification en terme d'AUC (voir figure 2.12C). Cette classification consiste à déterminer, à partir de la composition du microbiote, le statut clinique des patients (sain ou malade dans Qin-2014, et atteinte modérée ou sévère dans Loomba-2017). Nous avons pour cela utilisé des forêts aléatoires, qui utilisent les abondances relatives des espèces, ainsi que la richesse observée et la diversité de Shannon, après avoir réalisé une étape de sélection de variables, comme décrite

6. c'est la moyenne géométrique pour les valeurs non-nulles qui est utilisée ici

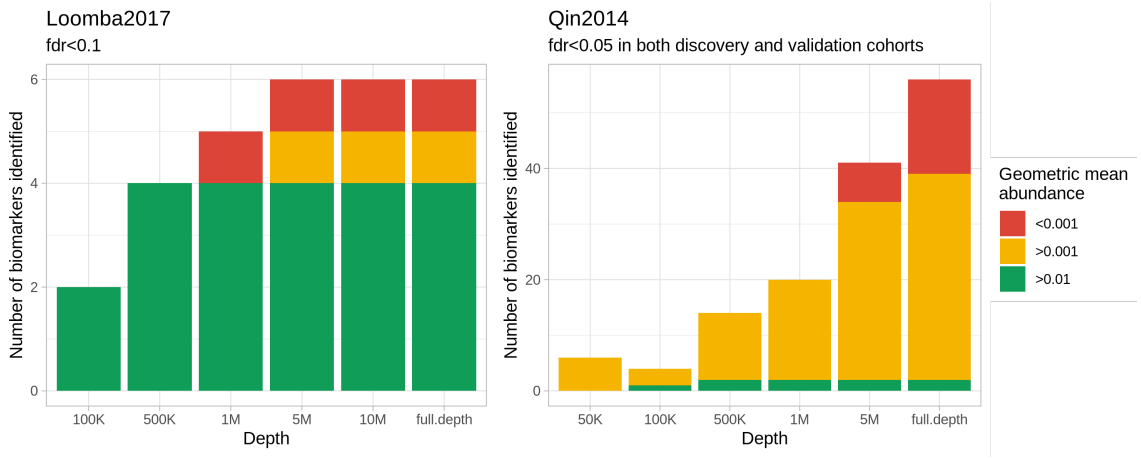


FIGURE 2.13 – Effet de la profondeur de séquençage sur la découverte de biomarqueurs entre les groupes de patients d'intérêt de Loomba-2017 et Qin-2014. La hauteur des barres représente le nombre total de biomarqueurs identifiés, qui sont décomposés selon leur abondance relative moyenne (moyenne géométrique sur les valeurs non-nulles).

par Loomba *et al.* [95]. Brièvement, cette méthode consiste à entraîner 300 forêts aléatoires, puis à utiliser l'ordre d'importance des variables données par la meilleure des 300 forêts pour éliminer itérativement chaque variable par ordre d'importance décroissante, jusqu'à obtenir le meilleur modèle. Pour le jeu de données de Loomba-2017, nous avons pris en compte l'âge et l'IMC comme variables non taxonomiques[7]. Concernant le jeu de données de Qin-2014, nous avons réalisé une première classification en utilisant tous les patients disponibles, puis une deuxième en entraînant le modèle sur la cohorte de découverte et en l'évaluant sur la cohorte de validation.
On observe sur la figure 2.12C que sur Loomba-2017, l'AUC décroît graduellement à partir de 1M *reads* par échantillon. Sur Qin-2014, la classification est très bonne, tant sur la cohorte de découverte que sur celle de validation. L'AUC sur la cohorte de validation a un léger décrochage entre 1 et 5M de *reads* par échantillon, mais la performance reste bonne (> 0.8) jusqu'à 500K *reads* par échantillon.

Résultats obtenus avec le filtre basique Les résultats présentés jusque là sont ceux obtenus avec le filtre utilisant les forêts aléatoires. En comparant ces résultats avec ceux obtenus avec le filtre basique, on note que les conclusions relatives à l'impact de la profondeur de séquençage sont les mêmes : il y a une bonne conservation du signal au moins jusqu'à 500K *reads* par échantillon, à part pour la classification de patients dans Loomba-2017 où la performance se dégrade à partir de 1M *reads* par échantillon. Cela montre que l'impact de la profondeur de séquençage sur la discrimination entre les groupes de patients est relativement faible, quelle que soit la stratégie de filtre utilisée.
Lorsqu'on compare les résultats obtenus par les deux types de filtres, on observe qu'il n'y a pas un filtre qui donne systématiquement des résultats plus significatifs que l'autre. Par exemple, la différence d'α-diversité entre les groupes dans Qin-2017 est plus significative avec le filtre basé sur les forêts aléatoires que le filtre basique (*cf.* figure 2.14A) alors que c'est l'inverse pour les différences de β-diversité dans Loomba-2017 (*cf.* figure 2.14B). Concernant la classification

7. Données récupérées en contactant les auteurs

de patients, le filtre basique donne souvent de meilleurs résultats (AUC plus élevée) que le filtre basé sur les forêts aléatoires (*cf.* figure 2.14C). Il semblerait que la classification soit moins impactée par le bruit introduit par les faux positifs présents dans les profils taxonomiques résultants du filtre basique, que par la perte de certaines populations clés lorsque l'on filtre pour assurer la qualité des profils. Cette robustesse vient sans doute de la propriété de sélection de variable des forêts aléatoires, qui lui permet de passer outre les nombreux faux positifs. En revanche, lors de l'évaluation de la classification sur la cohorte de validation de Qin-2014, les résultats obtenus à partir du filtre par forêt aléatoire sont meilleurs, ce qui conforte l'idée qu'ils produisent des données moins bruitées et ont une meilleure capacité de généralisation.

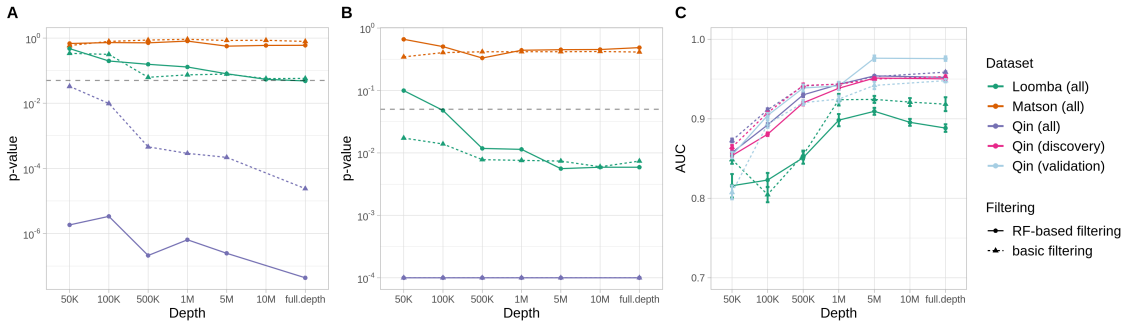


FIGURE 2.14 – Impacts de la stratégie de filtrage et de la profondeur de séquençage sur les différences entre les groupes de patients dans les différents jeux de données : significativité de la différence inter-groupes en termes (A) d'α-diversité (indice de Shannon, test de Wilcoxon) et (B) de β-diversité (PERMANOVA sur la matrice de distances de Bray-Curtis). Performance de classification (C) des patients selon leur groupe (classification par forêts aléatoires basée sur la composition de leur microbiote, *cf.* paragraphe 2.3.4 pour les détails). Les barres d'erreurs représentent l'écart-type de l'AUC à travers 10 répétitions du procédé de classification.

2.4 . Discussion

2.4.1 . Bilan

Nos simulations ont montré le besoin primordial de filtrer les données pour assurer la qualité des profils taxonomiques reconstruits. Nous avons vu que se contenter d'appliquer un seuil sur le nombre de *reads* et/ou sur la couverture des génomes était sous-optimal, et avons développé une approche basée sur un classifieur entraîné pour détecter les espèces réellement présentes et éliminer les faux positifs. Ces faux positifs étaient parfois identifiés par de très nombreux *reads* mais la méthode que nous avons développée prend en compte la répartition de ces *reads* le long du génome pour produire une classification fine. En fixant un taux maximal toléré de faux positifs dans les profils, nous avons pu comparer les profils obtenus, à travers les méthodes de filtres et les profondeurs de séquençage. Nous avons alors constaté que les filtres basés sur des classifieurs permettent d'identifier plus d'espèces que ceux basés sur des seuils. Nous avons également constaté que les profils étaient restreints aux espèces dominantes à faible profondeur de séquençage. Par exemple, à 500K *reads* par échantillon, les profils comptaient ~ 50 espèces par échantillon et étaient relativement complets pour les espèces présentes à une abondance relative supérieure à 4.10^{-4}. Comparativement, à 5M *reads* par échantillon, les profils comptaient ~ 100 espèces par échantillon et étaient fiables pour les espèces présentes à une abondance relative supérieure à 3.10^{-5}. On voit ici un résultat auquel on pouvait s'attendre : en séquençant 10 fois plus de *reads* par échantillon, on est capable d'observer des espèces 10 fois plus rares. Dans un certain nombre de contextes biologiques, n'identifier que les espèces les plus abondantes peut être suffisant.

Sur les données réelles, nos résultats montrent que les différences observées entres les groupes de patients en séquençage profond persistent à faible profondeur. On a observé cependant lors de la recherche de biomarqueurs (espèces différentiellement abondantes), que la réduction de la profondeur de séquençage s'accompagnait d'une diminution importante du nombre de biomarqueurs retrouvés, en particulier pour les plus rares. En utilisant un filtre basique, classiquement utilisé dans la littérature, nous avons observé sensiblement le même effet de la profondeur de séquençage sur les résultats. Ceci témoigne du fait que ces résultats sont relatifs au signal contenu dans les données elles-mêmes, et non pas spécifiques au pipeline que nous avons développé. En comparant les résultats obtenus avec les deux filtres, nous avons constaté dans certains cas que le filtre basique, dont on sait qu'il produit des profils taxonomiques avec de nombreux faux positifs, donnait des résultats légèrement plus significatifs pour la différence entre les groupes de patients. Ceci témoigne du fait que, dans ces cas, les résultats sont moins impactés par le bruit introduit par les nombreux faux positifs que par l'élimination, par le filtre basé sur les forêts aléatoires, de certaines espèces clés pour la distinction entre groupes. Au delà de la seule performance mesurée par l'AUC, les classifications réalisées à partir des profils taxonomique obtenus par le filtre utilisant les forêts aléatoires sont réalisés à partir d'espèces moins nombreuses, mais plus fiables (moins de faux positifs), ce qui rend les résultats beaucoup plus facilement interprétable, notamment pour la liste des espèces instrumentales pour la classification .

Nos résultats valident le fait que la métagénomique *shotgun* peu profonde est une approche intéressante pour établir un profil taxonomique du microbiote intestinal humain et explorer les différences entre groupes de patients dans un contexte clinique. Cependant, nous avons montré clairement que si on veut s'assurer de la fiabilité des profils taxonomiques, les profils obtenus en métagénomique *shotgun* peu profonde sont limités et doivent se restreindre aux espèces les

plus abondantes.
Le problème des faux positifs et de la fiabilité des profils taxonomiques obtenus par métagénomique *shotgun* peu profonde n'avait pas été discuté dans les études précédemment publiées sur le sujet. Il est possible que ce problème soit particulièrement présent dans nos données car nous utilisons un catalogue qui contient des espèces non-cultivées dont les génomes représentatifs sont incomplets. Néanmoins, la démarche de maîtriser le taux de faux positifs dans les profils taxonomiques est essentielle, pour assurer l'interprétabilité et la reproductibilité des résultats obtenus.

2.4.2 . Limitations

Une grande partie du travail est basée sur l'analyse d'un jeu de données simulé. Comme dans toute simulation, nous avons essayé de reproduire les conditions réelles, tout en ignorant certaines de leur caractéristiques. J'ai utilisé les compositions de microbiote intestinal de 100 patients, espérant couvrir ainsi une diversité de profil assez importante mais il aurait été peut-être intéressant d'en inclure encore plus. La composition de ces profils avait elle-même été obtenue à partir de données de métagénomique *shotgun* analysées par *MetaPhlAn*, la richesse moyenne était de ~ 100 espèces par échantillon, avec des abondances relatives des espèces allant aussi bas que 10^{-6}. La complexité des échantillons simulés est certainement inférieure à ce qu'on peut trouver en situation réelle. Par ailleurs, les profils taxonomiques étaient initialement donnés en suivant la taxonomie du NCBI, et ne contenaient que des espèces cultivées. Comme expliqué dans la section 2.2.1, j'ai transposé ces profils pour remplacer les espèces du NCBI en espèces de UHGG. Cette transposition est approximative, lorsque l'espèce n'était pas présente et nommée à l'identique dans l'UHGG que dans NCBI, j'ai choisi une espèce taxonomiquement proche. Cela a aussi pour conséquence d'ajouter aux simulations des espèces non cultivées. Enfin, les *reads* simulés ne comprenaient pas d'erreur de séquençage, ce qui n'est pas représentatif d'une situation réelle. Ce dernier point ne devrait avoir que peu d'influence sur les résultats car les taux d'erreurs des séquenceurs haut-débit de *reads* courts (*Illumina*) sont extrêmement faibles ($< 10^{-3}$), et négligeables devant d'autres sources de bruit que comprennent les données, notamment devant la diversité entre les génomes d'une même espèce. En effet, comme montré dans la figure 2.3, le fait de choisir le génome représentatif pour simuler les *reads* ou un autre génome appartenant à la même espèce, introduit une variabilité très importante.

Sur données réelles, les résultats sont très cohérents entre les 3 jeux de données quant à la reconstruction des profils taxonomiques et à l'impact de la profondeur de séquençage sur ceux-ci. En revanche, on a constaté une différence entre les jeux de données quant à l'effet de la profondeur de séquençage sur les différences observées entre les groupes de patients. Dans Qin-2104, le signal était très fort et très conservé à faible profondeur de séquençage, car il était porté par des espèces abondantes, alors que dans Loomba-2017, le signal était plus altéré par la profondeur de séquençage, même s'il restait détectable en SSM. Plus généralement, l'effet de la profondeur de séquençage sur la stratification des patients dépend complètement du signal biologique contenu dans les données. Si le signal est porté par des populations rares, il sera perdu par construction à faible profondeur de séquençage.

Le pipeline que j'ai développé utilise des forêts aléatoires pour filtrer efficacement les résultats d'alignement en identifiant les espèces susceptibles d'être réellement présentes et les faux positifs. Ces filtres, reposant sur un apprentissage automatique à partir d'un jeu de données d'entraînement, doivent être utilisés avec précautions dans des conditions similaires à celles qui ont servi à l'entraînement. Ainsi, si la démarche est transposable à d'autres écosystèmes,

à d'autres catalogues de référence et/ou à d'autres outils d'alignement, les classifieurs pré-entraînés doivent eux être restreints à l'analyse de données de microbiote intestinal humain alignées sur UHGG, et provenant d'un séquençage similaire aux conditions d'entraînements, en termes de taille et de nombre de *reads*. Dans d'autres conditions ou avec un autre catalogue, il faudrait procéder à de nouvelles simulations pour entraîner un nouveau classifieur.

Plus généralement, la métagénomique *shotgun* à faible profondeur est une alternative clairement intéressante mais qui s'accompagne de contraintes. La première est inévitablement la dépendance à un catalogue de référence. Celui-ci a une importance cruciale sur les résultats. L'exhaustivité du catalogue vis-à-vis de l'écosystème étudié est très importante et va déterminer le taux de *reads* qui seront alignés. Si ce catalogue n'inclut que des séquences provenant d'organismes préalablement cultivés, la fraction non-cultivée de l'écosystème ne sera pas caractérisée. À l'inverse, si le catalogue inclut des séquences provenant de métagénomique, dont la qualité et la taxonomie associée peuvent être discutables, les profils taxonomiques résultant héritent des défauts potentiels du catalogue. Dans notre cas, nous avons utilisé UHGG, qui contient plus de 4000 espèces, dont 70% n'ont pas de représentant cultivé, donc pas de nom latin, et dont l'existence n'est pas validée biologiquement.

2.4.3 . Perspectives

Pour compléter ce travail, il pourrait être intéressant de comparer les résultats présentés précédemment à ceux obtenus par d'autres outils d'analyses. Santiago-Rodriguez *et al.* [150] ont comparé l'alignement sur génomes complets et l'alignement sur gènes marqueurs et ont conclu que l'alignement sur génomes complets, tel que nous l'avons réalisé était plus pertinent dans le contexte d'une faible profondeur de séquençage. J'ai comparé sur le jeu de données de Qin-2014 les résultats obtenus avec `kraken2` [189], qui utilise un alignement exact des différents *k-mers* qui composent les *reads*, en utilisant toujours UHGG comme catalogue de référence. J'ai également utilisé `bracken` [97], qui permet d'affiner les profils taxonomiques obtenus par `kraken2` grâce aux *reads* ambigus. En alignant les mêmes *reads* sur les mêmes génomes de référence, j'avais fort heureusement des résultats similaires à ceux obtenus par `bwa mem`, que nous utilisons dans notre pipeline. Les résultats présentaient notamment le même problème de faux positifs (plusieurs milliers d'espèces identifiées). En revanche, les fichiers de sortie de `kraken2` et de `bracken` ne permettent pas de récupérer la fraction couverte des génomes. Le filtre ne peut donc se faire qu'en se basant sur le nombre de *reads* ou l'abondance relative (et non la fraction de génome couvert), ce qui est sous-optimal, comme longuement discuté dans la section 2.3.2.

Il existe dans la littérature quelques données sur l'impact de la profondeur de séquençage sur les gènes ou famille de gènes identifiées dans les données de métagénomique [69, 150, 148]. En revanche, il pourrait être intéressant d'étudier en détail comme nous l'avons fait la fiabilité des résultats obtenus. En particulier, l'approche que nous avons développée pour optimiser les filtres pourrait être appliquée pour filtrer les gènes, ou les différentes entités métagénomiques telles que les MGSs. Au lieu de n'utiliser qu'une règle construite sur le nombre de *reads* et sur la fraction couverte, il y a toutes les raisons de croire qu'utiliser une approche par classification pourrait améliorer les résultats, de manière analogue à ce que nous avons constaté au niveau des génomes.

La pertinence de la métagénomique *shotgun* à faible profondeur pourrait également être étudiée dans le cadre d'autres types d'écosystèmes. La contrainte la plus importante est l'existence d'un catalogue représentatif de l'écosystème. Dès lors qu'un tel catalogue existe, les méthodes

que nous avons utilisées pour optimiser la qualité des profils taxonomiques à faible profondeur de séquençage peuvent s'appliquer. La profondeur de séquençage nécessaire dépendra de la richesse de l'écosystème et de la question biologique.

2.4.4 . Conclusions

Nos résultats ont mis en avant la nécessité de filtrer finement les profils taxonomiques résultants de l'alignement de *reads* de métagénomique *shotgun* à faible profondeur sur les génomes complets pour obtenir des résultats fiables et interprétables. Les profils résultants étaient fortement affectés par la profondeur de séquençage, avec des profils limités aux espèces les plus abondantes lorsque l'on séquence à faible profondeur. Malgré ça, les différences observées à pleine profondeur entre les groupes de patients étaient conservées à faible profondeur de séquençage. Cela confirme que cette approche est adaptée à l'analyse de la composition du microbiote intestinal humain. En particulier, choisir la métagénomique *shotgun* à faible profondeur plutôt que le métabarcoding peut permettre de s'affranchir des biais d'universalité des amorces et du nombre de copies inhérents au métabarcoding, d'avoir un profil taxonomique résolu jusqu'aux espèces, et d'avoir accès à des informations fonctionnelles, sans augmenter les coûts de séquençage. Par ailleurs, certaines études pour lesquelles une analyse par métagénomique *shotgun* est envisagée pourraient être menées à faible profondeur si la baisse de coûts subséquents permet d'analyser plus d'échantillons. A budget équivalent, cela peut permettre d'inclure plus de patients, d'avoir une dimension longitudinale (plusieurs prélèvements par patients), ou encore de confirmer les résultats par une cohorte de validation indépendante. Ce choix doit être cependant considéré avec précaution si l'analyse fonctionnelle est l'objectif principal de l'étude, ou si la question biologique nécessite d'observer les taxa rares.
Dans une perspective d'analyse du microbiote intestinal dans un parcours de soin, la métagénomique *shotgun* à faible profondeur peut être utilisée si le but est d'évaluer l'α-diversité et d'identifier les espèces majoritaires, ou encore dans un contexte de classification diagnostique si celle-ci repose sur des espèces abondantes.

3 - COMBINAISON DE GÈNES MARQUEURS EN MÉTABARCODING

3.1 . Contexte

Le métabarcoding est une approche très utilisée pour la caractérisation de tous types d'écosystèmes microbiens. Il a l'avantage d'être peu coûteux et facile à mettre en place, et de convenir aux écosystèmes fortement contaminés par l'ADN de l'hôte, comme c'est le cas du microbiote vaginal. Cependant, comme présenté dans la table 1.1 page 19, il présente des limitations importantes. Les profils taxonomiques obtenus à partir de métabarcoding présentent des biais dûs (i) au manque d'universalité du gène marqueur et des amorces PCR pour amplifier le marqueur, (ii) au nombre de copies du marqueur dans le génome, (iii) au manque de résolution nucléotidique entre espèces et (iv) au manque de complétude des bases de données associées. Ce projet s'intéresse à la combinaison de marqueurs, qui est une alternative potentiellement intéressante pour s'affranchir des limitations inhérentes au métabarcoding classique (mono-marqueur).

3.1.1 . Complémentarité des marqueurs

Concernant les études d'écosystèmes bactériens, le gène 16S est de loin le gène marqueur le plus utilisé. En plus d'être universel, ce gène a l'avantage d'être constitué d'une alternance de régions conservées, permettant de concevoir des amorces universelles, et de régions peu conservées, dites "hypervariables", permettant de discriminer les espèces bactériennes. De plus, les bases de données de gènes 16S, comme SILVA [142], greengenes [176] ou encore RDP [29], sont très complètes. En revanche, le nombre de copies du gène 16S est variable d'un bactérie à l'autre (1 à 15 copies par génome [8, 169]) introduisant un biais du nombre de copies important dans les résultats [47, 81, 183]. Les régions hypervariables présentent également des différences en termes de résolutions taxonomiques, de biais d'amplification et de susceptibilité à la création de chimères. Le choix de la région doit donc être adaptée pour chaque écosystème.

D'autres gènes sont également utilisés, bien que plus rarement, et certains d'entre eux sont spécifiques à un écosystème donné. On trouve dans la littérature de nombreux d'articles qui mettent en avant les qualités de certains marqueurs. Comparés au 16S, ces marqueurs alternatifs apportent généralement une information plus précise en terme d'assignation taxonomique et/ou de quantification au prix d'une moindre universalité, puisqu'ils ne sont souvent pertinents que chez certains groupes de bactéries. Par exemple, chez les bactéries, le gène gyrB a permis une assignation taxonomique plus précise pour des microbiotes colonisant les aliments [137]. Le gène rpoB s'est avéré plus précis que le gène 16S pour étudier le microbiote dans les champs pétroliers [182] ainsi que des communautés bactériennes synthétiques associées à un nématode [125]. Le gène cpn60 a également apporté une information complémentaire et utile pour décrire le microbiote de champs pétrolier [84]. Le gène pheS a quant à lui permis une meilleure identification des *Lactobacillus* [162].

Il existe ainsi de nombreux cas où des marqueurs alternatifs apportent une information pertinente et complémentaire à celle apportée par le gène 16S, qui offre lui une description globale des écosystèmes bactériens.

3.1.2 . Intégration des résultats obtenus par plusieurs marqueurs

Combinaison de gènes marqueurs Au delà de la mise en évidence de la complémentarité des gènes marqueurs, on trouve également dans la littérature des études qui utilisent une approche multi-marqueurs. Ces études ont, pour chaque échantillon, réalisé plusieurs séquençages amplicon, ciblant différents gènes marqueurs.

La plupart de ces études utilisent plusieurs marqueurs pour élargir la couverture taxonomique de leurs analyses, car il n'existe pas de marqueur dont les amorces amplifieraient tous les

organismes d'intérêt de l'étude, à la fois procaryotes et eucaryotes. Frigerio *et al.* [53] ont combiné ITS2 et psbA-trnH pour identifier un maximum de plantes dans le but d'évaluer la composition des tisanes, alors que de Groot *et al.* [38] ont combiné les gènes 16S, 18S et rbcL pour identifier bactéries, champignons, protistes, animaux et plantes.

D'autres études qui s'intéressaient respectivement aux tardigrades (oursons d'eau) [174], et aux déjections des oiseaux pour en analyser le régime alimentaire [35], ont mis en avant l'intérêt d'avoir plusieurs marqueurs pour obtenir une meilleure résolution taxonomique et une plus grande confiance en les résultats lorsque les marqueurs identifiaient des taxa en commun. Da Silva *et al.* ont également développé et mis à disposition un programme dans un article subséquent [36] s'intéressant à la combinaison de marqueurs, mais leur démarche se limite à l'identification des taxa et ne cherche pas à améliorer l'estimation des profils d'abondances relatives.

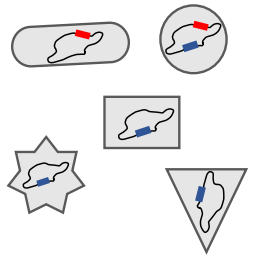
Enfin, Stefanni *et al.* [168] ont développé, dans le cadre de l'étude du zoo-plancton en mer Méditerranée, une approche visant à réellement combiner les résultats issus de métabarcoding ciblant les gènes 18S et COI, afin d'obtenir un profil taxonomique consensus, avec identification et estimation de leurs abondances relatives. Pour chaque marqueur, ils ont construit un profil taxonomique, en veillant à adapter le seuil de similarité de séquences au marqueur, puis retiré les taxa considérés comme de probables faux positifs selon les connaissances expertes sur l'écosystème. Les nombres totaux de *reads* des deux tables de comptages étant différents, ils ont appliqué un coefficient multiplicateur à l'une d'elle pour égaliser les sommes des comptages des deux tables[1]. Enfin, ils ont fusionné les deux tables de comptages comme suit : ils ont priorisé le comptage donné par COI (considéré comme plus résolutif) pour les taxa identifiés à la fois par COI et 18S, et conservé à l'identique le comptage des autres taxa, identifiés par un seul des deux marqueurs. Cette méthode présente plusieurs faiblesses. Mélanger les comptages après application du coefficient de normalisation n'est valide que si l'universalité des deux marqueurs est équivalente. En effet, comme montré dans la figure 3.1, si un premier marqueur n'est amplifié que chez une petite partie des organismes présents alors que l'autre est amplifié chez une grande partie, ramener les profils issus des deux marqueurs au même comptage total va grandement surestimer les organismes identifiés par le premier marqueur. De plus, il n'est pas clairement explicité comment les auteurs ont réalisé la fusion manuelle des tables de comptages, pour les taxas étant assignés à des niveaux taxonomiques différents entre les deux marqueurs. Il est ainsi difficile d'évaluer s'ils ont réellement pu tirer parti du pouvoir discriminant de chacun des marqueurs.

Les autres méthodes identifiées dans la littérature pour combiner plusieurs profils taxonomiques obtenus avec des marqueurs différents, sont dédiées à la combinaison de différentes régions hypervariables du gène 16S et sont détaillées ci-dessous.

Combinaison de différentes régions du gène 16S Fuks *et al.* [54] ont publié *SMURF* en 2018, une méthode qui a récemment été ré-implémentée dans un module de la plateforme qiime2 dénommée *Sidle* [40], et Schriefer *et al.* [155] ont publié *MVRSION* en 2018.

Comme illustré dans la figure 3.2, la première méthode s'appuie sur une base de données de référence (Greengenes v13.8 dans *SMURF* et Greengenes v13.8 et Silva v128 dans *Sidle*) et sur les paires d'amorces utilisées pour construire des séquences 16S complètes, ou partielles si

1. Il aurait été préférable de réaliser cette étape de normalisation à l'échelle de l'échantillon plutôt qu'à celle du jeu de données complet, de manière à tenir compte des écarts de profondeur de séquençage entre les échantillons



	Profil marqueur rouge	Profil marqueur bleu	Consensus Stefanni *et al.*
○	500	250	**500*** ou 250
▭	500	0	**500***
▭	0	250	250
✶	0	250	250
▽	0	250	250

FIGURE 3.1 – Représentation schématique d'un cas problématique pour la méthode de combinaison de profils multi-marqueurs développée par Stefanni *et al.* [168]. Le fait que les deux marqueurs (rouge et bleu) n'aient pas la même universalité, et qu'ils soient tous deux ramenés au même nombre de *reads* total (ici 1000 *reads* pour chaque marqueur) conduit à une surestimation de l'abondance relative de certaines bactéries, marquées en gras et avec une astérisque, et à la sous-estimation de toutes les autres. Concernant la bactérie circulaire, la valeur dans le profil consensus dépend du marqueur défini comme prioritaire.

toutes les régions ne sont pas ciblées, fruits de la concaténation des séquences produites par les différentes régions. En prenant en compte les probabilités d'erreur de séquençage et l'universalité des amorces, la méthode cherche à identifier quelles séquences de la base de données de référence sont présentes. Les abondances relatives de ces séquences sont ensuite estimées en maximisant la vraisemblance des données. Le résultat de la méthode est un ensemble de séquences concaténées et leurs abondances relatives. Ces séquences concaténées correspondent à des groupes d'entrées de la base de données partageant la même séquence nucléotidique sur les régions amplifiées. On peut alors assigner taxonomiquement ce groupe au dernier ancêtre commun des entrées formant le groupe, ou alors utiliser la séquence nucléotidique pour l'aligner sur une autre base de données (potentiellement plus à jour). Cette méthode n'est valide que si tous les organismes de l'écosystème étudié sont représentés dans la base de données de référence, car son résultat est l'abondance relative des séquences de la base de données.

La deuxième méthode, *MVRSION*, a pour objectif d'identifier les espèces et leurs abondances relatives, à partir du séquençage de 14 régions. La méthode, illustrée dans la figure 3.3, réduit la base de données de référence (SILVA v123) aux espèces candidates, qui ont été identifiées par l'alignement d'au moins 4 régions. Ensuite, en utilisant la base de données de référence réduite et les paires d'amorces utilisées, la méthode construit une liste de régions discriminantes pour chaque espèce. Une espèce est considérée comme effectivement présente si ses régions discriminantes sont suffisamment couvertes. La logique derrière cette méthode est de considérer qu'une espèce est présente si elle est retrouvée lors de l'alignement (même en cas d'alignement à plusieurs séquences de référence) ce qui crée de nombreux faux positifs, puis d'éliminer ces faux positifs s'ils ne sont pas retrouvés par les régions qui identifient cette espèce de manière unique. Cette méthode est elle aussi très dépendante de la base de données de référence qui est utilisée, et ne peut rendre qu'un résultat au niveau de l'espèce. En d'autres termes, elle ne permet pas de conserver l'ambiguïté de reconstruction et d'assigner à un niveau taxonomique supérieur si l'espèce observée n'est pas dans la base de données.

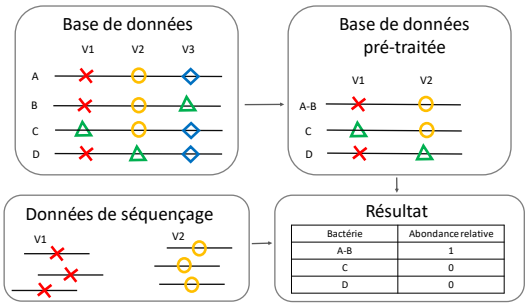


FIGURE 3.2 – Représentation schématique du concept de SMURF dans le cas où seules les régions V1 et V2 sont séquencées. Les séquences A et B de la base de données sont identiques sur V1 et V2, les deux séquences sont donc regroupées. Les données de séquençage permettent à l'algorithme de définir que c'est la séquence [croix-rond] qui est présente. Pour l'assignation taxonomique, on peut soit utiliser le dernier ancêtre commun de A et B de la base de données, soit aligner la séquence [croix-rond] sur une autre base de données (plus à jour).

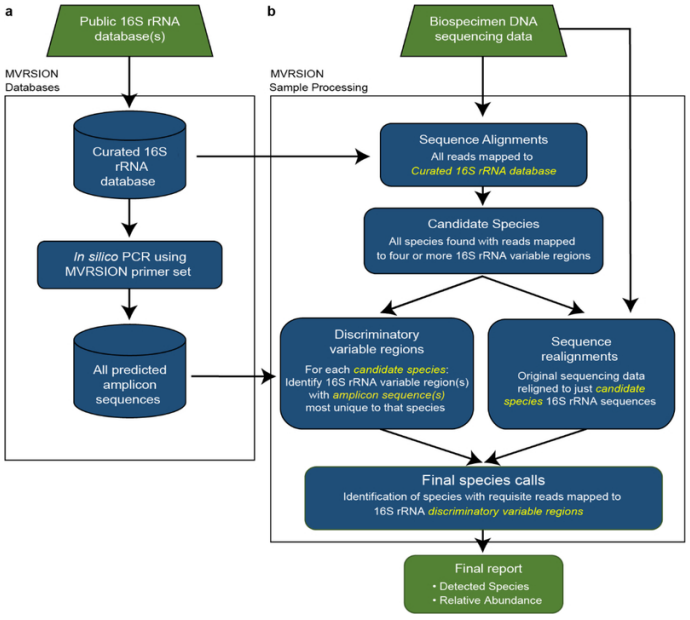


FIGURE 3.3 – Représentation schématique du fonctionnement de MVRSION, repris de [155]

Ces méthodes présentent certaines contraintes d'utilisation, notamment des licences logicielles payantes (*SMURF* est développé en Matlab et *MVRSION* requiert une licence *NovoAlign*), et imposent à l'utilisateur une base de données qui n'est pas à jour, en particulier Greengenes v13.8 qui a presque 10 ans. Elles ne sont adaptées que si tous les organismes de l'écosystème sont représentés dans la base de données (*SMURF* cherche les séquences 16S complètes et *MVRSION* rend un profil uniquement au niveau des espèces). De plus, elles ne sont applicables que pour le gène 16S et ne sont pas aisément transposables à une approche qui combine le gène 16S, et d'autres gènes marqueurs. En effet, *SMURF* requiert une base de données qui lie au sein d'une même séquence les différents amplicons. Cela pourrait théoriquement être contourné en utilisant comme référence une base de données de génomes complets, mais ces bases de données sont bien moins fournies que les bases de données de gènes marqueurs, en conséquence de quoi on perdrait beaucoup en représentativité de l'écosystème dans la base de données. La logique de *MVRSION* serait potentiellement applicable à d'autres gènes, mais une ré-implémentation serait nécessaire. De plus sa logique n'est pertinente qu'en cas de grand nombre de marqueurs avec chacun un faible pouvoir discriminant. Le nombre de copies des gènes, qui peut varier entre les marqueurs, constituerait potentiellement un point de blocage pour ces méthodes, qui ne prennent pas en compte cette problématique, car elle n'est pas pertinente lorsqu'on combine plusieurs régions du gène 16S.

3.1.3 . Objectifs

Compte-tenu de la forte complémentarité entre les marqueurs, et de l'absence de méthode satisfaisante permettant de combiner différents gènes marqueurs pour obtenir un profil taxonomique consensus, j'ai développé une méthode répondant à cette problématique. L'objectif est de combiner les profils taxonomiques issus de métabarcoding obtenus par différents marqueurs (qu'ils proviennent de plusieurs gènes ou de différentes régions du même gène) pour obtenir un profil consensus qui tire parti du pouvoir discriminant de chaque marqueur ainsi que de l'universalité de leurs amorces et de la complétude variable des bases de données. Le but est aussi bien d'affiner l'assignation taxonomique que d'améliorer l'estimation des abondances relatives. Pour convenir à un maximum d'utilisateurs, la méthode se doit d'être applicable à n'importe quel marqueur et n'importe quel écosystème. En ce sens, le mode de fonctionnement offrant le plus de liberté a été de laisser à l'utilisateur la responsabilité de construire un profil taxonomique par marqueur, éventuellement en utilisant des outils distincts, avant de lui proposer une méthode générique qui combine ces différents profils taxonomiques pour établir un profil taxonomique consensus, comme illustré schématiquement dans la figure 3.4.

Entrées et sorties de la méthode Concrètement, nous avons donc choisi de développer une méthode en `R`, qui prend en entrée plusieurs profils taxonomiques, correspondant à différents marqueurs, et calcule un profil taxonomique consensus. Un profil taxonomique correspond à une table de comptages, avec en ligne les entités taxonomiques choisies par l'utilisateur (ASVs, OTUs, ou comptages agglomérés à un rang taxonomique donné) et en colonne les échantillons, ainsi que la taxonomie associée à chaque entité. Nous avons pour l'instant utilisé comme entrées et sortie des objets de classe `phyloseq`, tels que défini dans le package *phyloseq* [107], car il est couramment utilisé par la communauté et on peut facilement créer un objet *phyloseq* à partir des résultats de n'importe quel pipeline d'analyse de données de métabarcoding. Il est nécessaire que les différents objets *phyloseq* que l'on cherche à combiner partagent les mêmes échantillons. Il faut également qu'ils partagent le même référentiel taxonomique, c'est-à-dire les mêmes rangs taxonomiques et les mêmes noms de taxa.

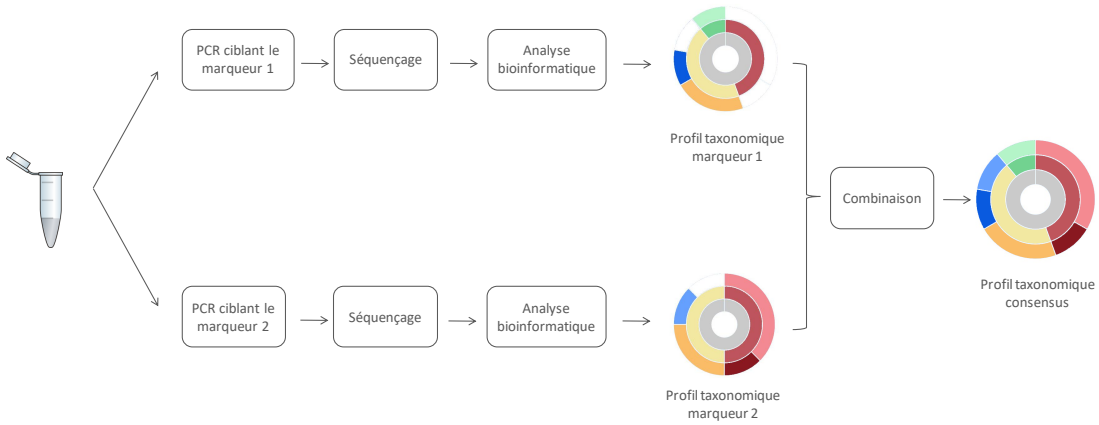


FIGURE 3.4 – Représentation schématique de la combinaison de deux profils taxonomiques en un profil consensus. Pour un échantillon dont on a préalablement extrait l'ADN, on réalise plusieurs PCRs pour amplifier plusieurs gènes marqueurs, qui sont séquencés puis analysés par un pipeline bioinformatique. Cette analyse bioinformatique comprend classiquement un prétraitement, un débruitage (ASVs) ou *clustering* (OTUs) puis un alignement sur une base de données de référence pour assigner taxonomiquement les séquences, et donne en sortie un profil taxonomique (identification des taxa, de leurs affiliations taxonomiques et de leurs abondances relatives). Les profils issus de chaque marqueur sont ensuite combinés, par la méthode que j'ai développée, pour donner un profil consensus.

Scénarios d'utilisation Au-delà de la souplesse dans les choix d'outils et bases de données de référence utilisés pour chaque marqueur, nous avons voulu développer une méthode qui puisse intégrer des résultats obtenus avec des marqueurs ayant différentes couvertures taxonomiques, certains universels et d'autres plus spécifiques à une clade par exemple, et séquencés à différentes profondeurs. Par exemple, il pourrait être intéressant dans un grand nombre d'applications biologiques d'utiliser un marqueur principal, très universel et séquencé avec une grande profondeur, de manière à avoir une vue d'ensemble de l'écosystème, et un ou plusieurs séquençages "satellites" ciblant d'autres gènes marqueurs, qui auraient pour objectif de raffiner le profil taxonomique dans certaines clades d'intérêt pour lesquelles le premier marqueur n'était pas résolutif. On peut également envisager un scénario en deux temps pour certains projets : un premier séquençage préliminaire qui met en évidence l'importance de certaines clades mais manque de résolution, et un ou plusieurs séquençages ultérieurs ciblant un gène marqueur qui apporte une meilleure résolution pour ces clades, produisant ainsi une information complémentaire pour un coût modéré.

3.1.4 . Cas du microbiote vaginal

L'analyse du microbiote vaginal représente un cas d'application idéal de cette méthode et en est la motivation première puisqu'il constitue un sujet d'actualité pour mon équipe à Alphabio au moment de mon doctorat. Ainsi, la suite de ce projet va se concentrer sur l'analyse de cet écosystème.

Le microbiote vaginal se divise en 5 classes (ou CST pour *Community State Types*) [144]. Les CST I, II, III et V sont définies par une forte dominance de *Lactobacillus crispatus*, *L. gasseri*, *L. iners* et *L. jensenii* respectivement. La CST IV n'est au contraire dominée par aucune espèce du genre *Lactobacillus* et se caractérise à la fois par cette absence de dominance, par un pH

plus élevé dû notamment à l'absence d'acidification du milieu par des *Lactobacillus* et enfin par une plus grande biodiversité favorisée par le pH moins acide. La CST IV correspond à un état qui peut parfois être pathologique, on parle alors de vaginose bactérienne. La vaginose bactérienne est impliquée entre autres, dans les problèmes d'infertilité.

Techniques d'analyses utilisées en pratiques courantes En biologie clinique, l'analyse du microbiote vaginal est traditionnellement réalisé par observation microscopique des sécrétions vaginales. Le test de Nugent consiste à observer les bactéries après coloration de Gram, pour dénombrer 3 morphotypes bactériens reconnaissables par un biologiste médical : *Lactobacillus spp.*, *Gardnerella vaginalis* et *Mobiluncus spp.*. En établissant un rapport entre l'abondance des *Lactobacillus spp.* et celle des autres bactéries, le score de Nugent définit le diagnostic de la vaginose bactérienne.

La caractérisation du microbiote vaginal par qPCR est également possible, bien qu'elle ne soit pas utilisée en pratique clinique courante. Cette méthode vise à estimer la concentration bactérienne d'une espèce, en utilisant des amorces qui lui sont spécifiques. Il est possible de faire plusieurs qPCR pour évaluer la concentration de plusieurs espèces, et comparer les résultats pour évaluer laquelle est dominante. La région ciblée par les amorces étant souvent une portion du gène 16S, les résultats souffrent du biais précédemment évoqué lié au nombre de copies. Cette méthode est ciblée sur quelques espèces préalablement choisies, et ne rend pas compte de l'écosystème dans son ensemble. Elle peut donner des résultats erronés si des bactéries abondantes dans l'écosystème ne sont donc pas ciblées.

Métagénomique Le séquençage à haut débit est bien évidemment très utilisé, surtout dans la recherche, mais n'est pas encore couramment utilisé en pratique clinique. Le séquençage *shotgun* est relativement peu utilisé, comparativement au microbiote intestinal notamment, principalement à cause du fort taux d'ADN humain retrouvé dans un prélèvement vaginal. En effet, entre 90 et 99% des *reads* obtenus par séquençage shotgun sont de l'ADN humain (*cf.* figure 1.4), et sont donc retirés si on s'intéresse au microbiote, rendant les coûts de séquençage particulièrement élevés relativement à l'information produite. S'il existe des méthodes biochimiques pour réduire le taux d'ADN humain, elles introduisent des biais non-négligeables dans l'estimation des abondances relatives des espèces[1, 102].

Métabarcoding Le métabarcoding, qui offre une vue d'ensemble de la composition du microbiote vaginal et ne souffre pas du problème lié au fort taux d'ADN de l'hôte, représente une alternative pertinente et très utilisée. La majorité des études sont réalisées en ciblant la région 16S, et la question du choix de la région 16S la plus appropriée pour le microbiote vaginal a été amplement discutée [164, 62, 178]. Un enjeu important consiste à identifier les espèces du genre *Lactobacillus* et ainsi pouvoir déterminer la classe d'appartenance du microbiote vaginal. Le gène 16S étant très conservé au sein du genre *Lactobacillus*, la discrimination entre les espèces est difficile. Il arrive d'avoir deux espèces partageant strictement la même séquence selon certaines régions, c'est le cas par exemple de *L. gasseri* et *L. paragasseri*, ou encore *L. jensenii* et *L. mulieris* sur les régions V1V4 et V3V4. Les espèces dont les séquences amplifiées sont extrêmement proches seraient regroupées dans la même OTU, et c'est d'ailleurs un des exemples qui a été utilisé pour plébisciter le concept d'ASV [25]. Lors de l'assignation taxonomique, la discrimination entre espèces, lorsqu'elle est possible, se base parfois sur 1 ou 2 nucléotides, ce qui peut conduire a des erreurs de classification. En cas de stricte égalité entre

plusieurs *hits*, certains utilisent le fait que des espèces ne sont pas connues pour être présentes dans le microbiote vaginal pour trancher et identifier l'espèce, ce qui peut poser problème dans les cas particuliers où l'espèce de *Lactobacillus* n'est pas l'une des quatre majoritaires. En effet, même si les quatre espèces majoritaires de *Lactobacillus* qui définissent les CSTs dominent le microbiote vaginal, jusqu'à 32 autres espèces de *Lactobacillus* ont déjà été retrouvées [41] dans cet écosystème. Par ailleurs, les différentes copies du gènes 16S ne sont pas toujours identiques et la distance nucléotidique entre deux copies peut parfois se confondre, voire dépasser, celles entre deux espèces.

Il n'existe donc pas de méthode qui permette de caractériser le microbiote vaginal à faible coût, à haut débit, de manière exhaustive (non limitée à quelques espèces ou morphotypes), et qui soit suffisamment résolutive pour identifier de manière fiable les espèces de *Lactobacillus* et ainsi les CSTs. La combinaison de marqueurs en métabarcoding constitue donc une alternative potentiellement d'intérêt pour répondre à toutes ces problématiques.

3.2 . Travaux préliminaires

Pour simplifier la lecture du manuscrit, on appellera "marqueur" la région ciblée lors d'une expérience de métabarcoding. Chaque marqueur est défini par la paire d'amorces qui sert à amplifier la région ciblée lors de la PCR. Un marqueur peut donc être un gène entier, une partie d'un gène, ou une région intergénique.

Objectifs Le but de ce travail préliminaire est de déterminer les marqueurs les plus pertinents pour la caractérisation du microbiote vaginal.

Pour faciliter le travail, nous nous sommes restreints à des marqueurs déjà utilisés en métabarcoding. D'après la littérature, j'ai établi une liste de gènes marqueurs et paires d'amorces comme montré dans la table 3.1. Cette liste n'est pas exhaustive mais contient l'ensemble des marqueurs candidats qui seront évalués sur la base des caractéristiques suivantes :

- l'universalité des amorces pour les bactéries d'intérêt ;
- le nombre de copies amplifiées au sein de chaque génome ;
- la taille de la région amplifiée, et ainsi la possibilité d'assembler la paire de *reads* de séquençage grâce à leur chevauchement ;
- la diversité intra- et inter-spécifique au sein de la région amplifiée, et ainsi le pouvoir discriminant du marqueur, particulièrement chez les espèces du genre *Lactobacillus* ;
- la complétude de la base de données de référence associée au marqueur.

Pour certains gènes, j'ai considéré plusieurs paires d'amorces lorsque les données de la littérature semblaient pertinentes, d'où les versions notées "marqueur$_{bis}$".

Génomes d'intérêt Une liste de 18 espèces de bactéries d'importance clinique a été établie par Dr. Marion Bonnet, pharmacienne et microbiologiste en charge du microbiote vaginal à Alphabio. Cette liste, détaillée dans la table 3.2, comporte les 4 espèces prépondérantes du genre *Lactobacillus* ainsi que des bactéries fréquemment présentes en cas de vaginose bactérienne.

Pour chacune de ces espèces, j'ai récupéré les génomes disponibles dans la base de données RefSeq via l'outil `ncbi-genome-download`. J'ai d'abord priorisé les génomes complets, puis les assemblages au niveau chromosomique et enfin les assemblages incomplets lorsque seuls ceux-ci étaient disponibles. J'ai ainsi obtenu 94 génomes. Dans le but de pourvoir évaluer, pour chaque

Marqueur	Ref.	Taille de l'amplicon	Séquence sens (5'-3')	Séquence anti-sens (5'-3')
16S V1V4		750-900	GAGTTTGATCMTGGCTCAG	CTACCAGGGTATCTAATCC
16S V3V4		400-500	CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC
gyrB	[137]	250-350	MGNCCNGSNATGTAYATHGG	CNCCRTGNARDCCDCCNGA
pheS	[116]	400-500	CAYCCNGCHCGYGAYATGC	CCWARVCCRAARGCAAARCC
$pheS_{bis}$	[116, 162]	400-500	CAYCCNGCHCGYGAYATGC	GGRTGRACCATVCCNGCHCC
$cpn60_1$	[152]	550-650	GAIIIIGCIGGIGAYGGIACIACIAC	YKIYKITCICCRAAICCIGGIGCYTT
$cpn60_2$	[152]	550-650	GAIIIIGCIGGYGACGGYACSACSAC	CGRCGRTCRCCGAAGCCSGGIGCCTT
$cpn60_{bis}$	[193]	550-650	GCYGGTGCWAACCCNGTTGG	AANGTNCCVCGVATCTTGTT
ITS_{bact}	[110]	700-1500	KRGGRYKAAGTCGTAACAAG	TTTTCRYCTTTCCCTCACGG
$ITS_{bact,bis}$	[114]	700-1500	GGGCTACACACGYGCWAC	GCCWAGGCATCCDCC
rpoA	[116]	450-650	ATGATYGARTTTGAAAAACC	ACHGTRTTRATDCCDGCRCG
$rpoA_{bis}$	[115]	450-650	ATGATYGARTTTGAAAAACC	ACYTTVATCATNTCWGVYTC
rpoB	[125]	400-500	GGYTWYGAAGTNCGHGACGTDCA	TGACGYTGCATGTTBGMRCCCATMA

TABLE 3.1 – Liste des marqueurs candidats avec les références bibliographiques correspondantes, la taille approximative de la région amplifiée et les séquences nucléotidiques des amorces de PCR.

marqueur, la diversité au sein du genre *Lactobacillus*, j'ai téléchargé, selon le même mode de fonctionnement, 103 génomes appartenant aux 36 autres espèces du genre *Lactobacillus*.

Amplification J'ai aligné les séquences des amorces de PCR sur les 197 génomes avec `bowtie2` [86] en adaptant le système de score [2] pour permettre l'alignement de petites séquences (20-30 nucléotides) en tolérant un ou plusieurs *mismatchs* entre les amorces et le génome. Ces options sont utilisées pour gérer le cas où les séquences des amorces et du génome ne sont pas identiques mais suffisamment proches pour permettre l'amplification. Cependant, le nombre maximal de *mismatchs* au delà duquel il n'y a plus d'amplification n'est pas quantifiable de manière exacte et universelle : il dépend notamment de la position du *mismatch* dans l'alignement [129] et du contenu nucléotidique des amorces, qui contribue à définir les conditions expérimentales de PCR. De plus, l'amplification ou non d'une séquence n'est pas un phénomène binaire : le nombre de *mismatchs* va jouer dans l'affinité, et donc dans la sur-représentation ou sous-représentation des séquences dans le produit de PCR, en prolongement du biais d'amplification. Néanmoins, dans les résultats qui suivent, et par souci de simplicité, j'ai fixé la limite à 1 *mismatch*. C'est un seuil plutôt stringent de considérer que les séquences ne sont pas du tout amplifiées à partir de 2 *mismatchs*, mais les séquences seraient de toute manière sous-représentées dans les produits de PCR à cause de leur manque d'affinité avec les amorces. J'ai récupéré les coordonnées génomiques où étaient alignées les amorces sens et anti-sens avec `samtools`, et ainsi récupéré les régions amplifiées. J'ai retenu les régions qui avaient une taille conforme à ce qui est trouvé dans la littérature, tels que montré dans la table 3.1.

2. une graine de taille 10 avec un *mismatch* toléré dans la graine, pas d'insertion/délétion, pas de bonus pour un match, une pénalité de 1 pour un *mismatch* et un score minimal égal à l'opposé du nombre de *mismatch* qu'on souhaite tolérer `-L 10 -N 1 -rdg 5,5 -rfg 5,5 -ma 0 -mp 1,1 -np 0 -score-min C,-[nb.mismatch]`

Identifiant taxonomique	Nom de l'espèce
147802	*Lactobacillus iners*
1596	*Lactobacillus gasseri*
47770	*Lactobacillus crispatus*
109790	*Lactobacillus jensenii*
699240	BVAB1 (*Candidatus Lachnocurva vaginae*)
884684	BVAB3 (*Mageeibacillus indolicus*)
907	*Megasphaera elsdenii*
187101	*Sneathia vaginalis* (*Leptotrichia amnionii*)
40543	*Sneathia sanguinegens*
84112	*Eggerthella lenta*
2052	*Mobiluncus mulieris*
2051	*Mobiluncus curtisii*
2098	*Mycoplasma hominis*
82135	*Fannyhessea vaginae* (*Atopobium vaginae*)
2702	*Gardnerella vaginalis*
28125	*Prevotella bivia*
386414	*Prevotella timonensis*
2130	*Ureaplasma urealyticum*

TABLE 3.2 – Liste des espèces bactériennes d'intérêt clinique du microbiote vaginal utilisées pour l'analyse préliminaire. Les noms d'espèces entre parenthèses sont des synonymes taxonomiques.

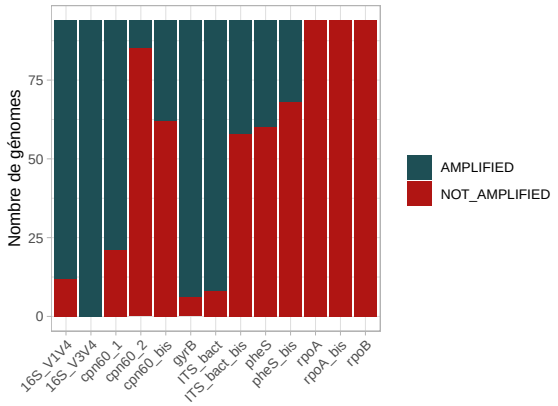


FIGURE 3.5 – Amplification des 94 génomes des espèces importantes du microbiote vaginal par les différentes paires d'amorces.

On observe sur la figure 3.5 que les marqueurs 16S V1V4 et V3V4, ITS_{bact}, gyrB et cpn60_1 sont les seuls à être amplifiés chez la très grande majorité des génomes considérés. Seules les amorces de 16S V3V4 sont parfaitement universelles pour l'ensemble des génomes considérés. Les amorces correspondant à rpoA et rpoB n'ayant rien amplifié, ces marqueurs seront écartés pour la suite.

En observant le détail des génomes amplifiés (table 3.3), on peut voir que les marqueurs cpn60_2, cpn60_{bis}, $\text{ITS}_{bact,bis}$ et pheS_{bis} ne sont amplifiés que chez une partie des génomes des 4 espèces prépondérantes de *Lactobacillus*, ce qui pourrait compromettre *in fine* l'identification

Nom du marqueur	*L. crispatus*	*L. gasseri*	*L. iners*	*L. jensenii*	Autres espèces du genre *Lactobacillus*	Autres espèces du microbiote vaginal
16S V1V4	16/16	9/9	7/7	2/2	93/103	48/60
16S V3V4	16/16	9/9	7/7	2/2	99/103	60/60
cpn60$_1$	16/16	9/9	7/7	2/2	103/103	39/60
cpn60$_2$	0/16	0/9	0/7	0/2	0/103	9/60
cpn60$_{bis}$	16/16	9/9	5/7	2/2	98/103	0/60
gyrB	16/16	9/9	7/7	2/2	103/103	54/60
ITS$_{bact}$	16/16	9/9	7/7	2/2	76/103	52/60
ITS$_{bact,bis}$	16/16	9/9	7/7	0/2	66/103	4/60
pheS	16/16	9/9	7/7	2/2	86/103	0/60
pheS$_{bis}$	16/16	9/9	0/7	1/2	67/103	0/60

TABLE 3.3 – Détail de l'amplification des génomes par les différentes paires d'amorces. Les marqueurs apparaissant en gras sont ceux qui ont été sélectionnés pour la suite de l'analyse. Pour chaque espèce (ou groupe d'espèces) et chaque marqueur, le tableau contient le nombre de génomes amplifiés sur le nombre total de génomes : par exemple le marqueur 16S V1V4 a été amplifié dans 48 génomes parmi les 60 génomes des autres espèces (*i.e.* hors *Lactobacillus*) du microbiote vaginal.

des CSTs. Ces marqueurs sont donc retirés pour la suite de l'analyse au profit des autres marqueurs qui n'ont pas ce défaut. On observe également que pheS est amplifié chez toutes les espèces d'intérêt de *Lactobacillus* mais chez aucune des autres espèces, contrairement aux autres marqueurs jusqu'alors sélectionnés.

Taille des amplicons La distance entre les positions où sont alignées les amorces sens et anti-sens, permet de déterminer la taille des amplicons qu'on obtiendrait en réalisant la PCR. Cette taille varie selon les marqueurs, et parfois entre les différents génomes pour un même marqueur. Il est important de la connaître pour déterminer les conditions expérimentales de la PCR (notamment pour ajuster le temps d'élongation) et pour savoir si les *reads* appariés seront chevauchant et s'il sera possible de les assembler. Si ce chevauchement est souvent recommandé [16], il n'est pas strictement nécessaire : moyennant une adaptation des *pipelines* bioinformatiques [14], on peut analyser les données et cela permet même de cibler des régions plus résolutives pour certains écosystèmes [134]. Ce ne sera donc pas un critère éliminatoire.

On observe sur la figure 3.6 que les *reads* produits par l'amplification des régions 16S V1V4, cpn60 et ITS$_{bact}$ ne pourront pas être assemblés (taille supérieure à 482 bp, la taille maximale de deux *reads* pairés (251 pb) chevauchants d'au moins 20 nucléotides), et que la taille du segment amplifié par les amorces de ITS$_{bact}$ est très variable selon les génomes, ce qui est conforme aux connaissances existantes sur cette région intergénique.

Nombre de copies Le nombre de copies du segment amplifié est une information cruciale dans la perspective de comparer puis d'intégrer des profils taxonomiques issus de différents marqueurs.

On observe dans la figure 3.7 et la table 3.4 que cpn60 et pheS ont la propriété d'être présents en un seul exemplaire dans les génomes bactériens considérés. Le nombre de copies obtenues avec les amorces gyrB varie entre 1 et 2 (il y a une exception à 3 copies), ce qui

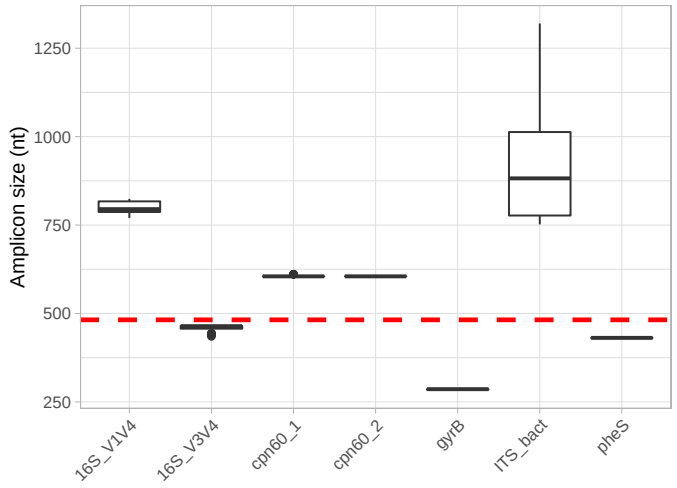


FIGURE 3.6 – Taille des amplicons obtenus à partir des 94 génomes d'intérêt pour les différents marqueurs. La ligne rouge représente la taille maximale d'un amplicon pouvant être intégralement séquencé et assemblé en Illumina (séquençage pairé MiSeq $2 * 251\ bp$, avec un chevauchement entre les *reads* d'au moins 20 nucléotides).

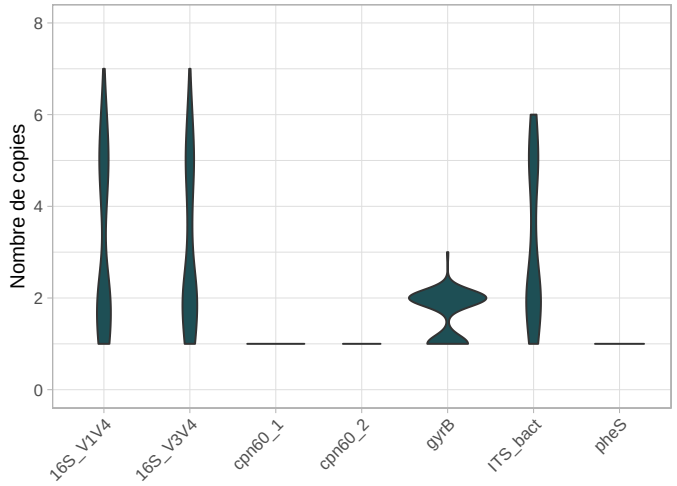


FIGURE 3.7 – Distribution du nombre de copies amplifiées par génomes, au sein des 94 génomes d'intérêt, pour chaque marqueur. Les génomes pour lesquels les marqueurs n'étaient pas amplifiés ont été préalablement retirés.

s'explique par l'amplification de parE, un gène paralogue à gyrB [137]. Ce paralogue n'est amplifié que chez certains génomes, et cause un biais en tout point similaire au biais du nombre de copies. On note sur la figure 3.7 et la table 3.4 que les amorces 16S V1V4, V3V4 et ITS_{bact} partagent le même biais du nombre de copies, ceci s'expliquant par leur juxtaposition au sein de l'opéron contenant l'ADN ribosomique. La figure 3.8 confirme cette information : on voit qu'une grande partie des génomes sont sur la diagonale. Quelques génomes sortent de la diagonale, traduisant le fait que, dans certains cas, seules certaines copies sont amplifiées. Ils ont entre 4 et 6 copies dans les génomes des *Lactobacillus* et en moyenne moins de 2 dans les autres génomes d'intérêt.

Nom du marqueur	*L. crispatus*	*L. gasseri*	*L. iners*	*L. jensenii*	Autres espèces du microbiote vaginal
16S_V1V4	4.9 ± 0.2	4.9 ± 0.9	5.9 ± 0.4	4 ± 0	1.8 ± 1
16S_V3V4	4.9 ± 0.2	4.9 ± 0.9	5.9 ± 0.4	4 ± 0	1.9 ± 1
cpn60_1	1 ± 0	1 ± 0	1 ± 0	1 ± 0	1 ± 0
cpn60_2					1 ± 0
gyrB	2 ± 0	2 ± 0	2 ± 0	2 ± 0	1.5 ± 0.5
ITS_bact	4.9 ± 0.2	5 ± 0.9	5.9 ± 0.4	4 ± 0	1.9 ± 0.6
pheS	1 ± 0	1 ± 0	1 ± 0	1 ± 0	

TABLE 3.4 – Détail du nombre de copies amplifiées par les différentes paires d'amorces : $moyenne \pm ecart\ type$.

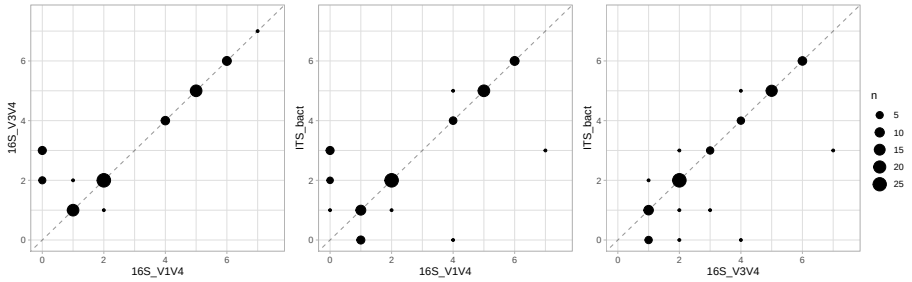


FIGURE 3.8 – Relation entre le nombre de copies amplifiées dans les 94 génomes d'intérêt pour les trois marqueurs appartenant à l'ADN ribosomal (16S V1V4, 16S V3V4 et ITS$_{bact}$). Les axes des abscisses et des ordonnées renseignent sur le nombre de copies pour chaque marqueur, et la taille des points représente le nombre de génomes.

Diversité entre les séquences amplifiées Pour chaque marqueur, afin d'étudier la diversité intra et inter-espèces, j'ai réalisé une comparaison deux à deux des séquences amplifiées chez l'ensemble des génomes d'intérêt, via un alignement global réalisé avec la fonction `pairwise2.align.globalxx` de la librairie `biopython`. Comme expliqué en introduction de ce chapitre, il est crucial d'identifier les espèces du genre *Lactobacillus*, ce que le seul séquençage 16S ne permet pas, ou uniquement de façon partielle. Dans le but d'avoir une première idée du pouvoir discriminant des différents marqueurs chez les *Lactobacillus*, j'ai étudié la similarité entre les séquences, exprimée en pourcentage d'identité, en fonction de leur relation taxonomique, en utilisant la même approche que Kim *et al.* [83]. J'ai aussi inclus dans ces comparaisons les copies multiples au sein d'un même génome si celles-ci n'étaient pas strictement identiques.

On observe dans la figure 3.9 que les résultats de ces comparaisons sont très différents selon les marqueurs. L'aspect primordial pour l'identification des *Lactobacillus* d'intérêt est que la similarité au sein d'une espèce (courbes jaunes) soit plus élevée que la similarité entre séquences du même genre et d'une espèce différente (courbes vertes). On observe pour les marqueurs 16S V1V4 et V3V4, que la densité n'est pas nulle à 100% d'identité au niveau du genre, témoignant du fait que certaines espèces distinctes du même genre peuvent avoir des séquences identiques, ce qui rend impossible l'identification de l'espèce dans ces cas-là. Pour tous les autres marqueurs, on observe une diversité inter-espèces plus importantes, témoignant ainsi d'un plus grand pouvoir discriminant. Concernant les amorces ciblant gyrB et ITS$_{bact}$, on observe des séquences appartenant à la même espèce voire au même génome qui ont des pourcentages

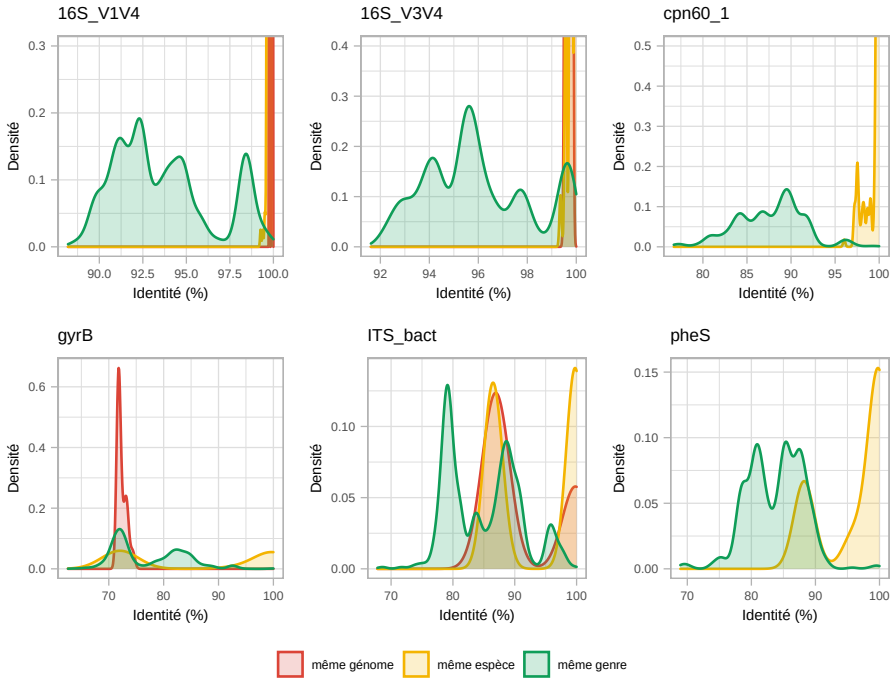


FIGURE 3.9 – Comparaison deux à deux des séquences amplifiées : pourcentage d'identité pour chaque paire de séquences en fonction de la relation entre ces deux séquences, pour les différents marqueurs. Les comparaisons représentées dans les courbes de densité comparent les séquences provenant des génomes des espèces d'intérêt de *Lactobacillus* aux séquences provenant de tous les génomes de *Lactobacillus*.

d'identités faibles, s'expliquant par une forte diversité entre les différentes séquences paralogues au sein des génomes. Concernant gyrB, le pic observé entre 70 et 75 % d'identité correspond à parE, le paralogue de gyrB qui est amplifié chez certains génomes, comme décrit dans Poirier *et al.* [137]. On note au contraire que les différentes copies du 16S ont une très forte similarité au sein d'un génome.

Bilan Basé sur l'ensemble de ces informations, j'ai choisi de retenir les marqueurs 16S V1V4 et V3V4, ITS_{bact}, gyrB et $cpn60_1$ pour la suite du projet. pheS est amplifié chez tous les *Lactobacillus* d'intérêt mais n'est amplifié chez aucune autre espèce d'intérêt. Il pourrait éventuellement être utilisé uniquement dans le but de discriminer les *Lactobacillus*, mais nous cherchons à ce stade des marqueurs plus universels, pour apporter un maximum d'information. Aussi, nous avons décidé d'écarter pheS, qui n'avait pas d'avantage comparativement aux autres marqueurs alternatifs.

Compte-tenu des observations faites lors de ce travail préliminaire, nous sommes en mesure de faire des conjectures sur les profils taxonomiques résultants des différents marqueurs retenus :

— les profils obtenus par 16S V3V4 captureront le mieux la biodiversité (les amorces permettent d'amplifier un plus grand nombre d'espèces), notamment pour la CST IV (figure 3.5, table 3.3) ;

— dans un échantillon qui comporte un mélange d'une espèce de *Lactobacillus* et d'autres espèces, typiquement les CST I, II, III et V, les marqueurs gyrB, ITS_{bact}, 16S V1V4 et $cpn60_1$ sont susceptibles de ne pas être amplifiés chez certaines des autres bactéries

que les *Lactobacillus*, celles-ci seraient donc absentes des profils taxonomique résultant, provoquant également une sur-estimation des *Lactobacillus* (figure 3.5, table 3.3) ;
— le biais du nombre de copies provoquera une sur-estimation des *Lactobacillus* dans les profils obtenus par les marqueurs 16S V1V4, V3V4 et ITS_{bact} (figure 3.7, table 3.4) ;
— les marqueurs ITS_{bact}, gyrB et $cpn60_1$ seront plus résolutifs pour discriminer les espèces de *Lactobacillus* et ainsi identifier les CST I, II , III et V (figure 3.9).

Par cette analyse préliminaire, j'ai pu montrer (i) qu'il n'y a pas de marqueur parfait pour caractériser le microbiote vaginal, (ii) que les différents marqueurs sont complémentaires, certains apportant les informations qui manquent aux autres, et donc (iii) qu'il serait très intéressant de développer une méthode qui combine les informations apportées par plusieurs marqueurs. Enfin, on peut aussi noter que le biais du nombre de copies, différent entre les marqueurs, sera source d'incohérence entre les profils obtenus par les différents marqueurs, ce qui constituera un défi au moment d'obtenir un consensus entre ceux-ci.

3.3 . Matériel et méthode

Comme évoqué précédemment (cf. section 3.1.3 page 56), la méthode que je me propose de développer permet de combiner plusieurs profils taxonomiques, laissant ainsi à l'utilisateur libre choix concernant les outils et bases de données de référence qui conviennent à ses données, à son écosystème d'intérêt, *etc.* Le paragraphe qui suit détaille les choix que nous avons faits pour mener le projet, mais la méthode qui permet d'obtenir le profil consensus, et qui sera décrite dans la section 3.3.2, est indépendante de ces choix-là.

3.3.1 . Développement du pipeline pour l'assignation taxonomique de chaque marqueur

Bases de données de référence Pour chaque marqueur il a fallu choisir une base de données, et la transposer dans un référentiel taxonomique commun. En effet, il est impératif que les bases de données qui sont utilisées pour réaliser l'assignation taxonomique partagent rigoureusement la même taxonomie. Il s'agit d'une condition nécessaire pour pouvoir *in fine* intégrer les profils taxonomiques et obtenir un consensus. Pour ce faire, j'ai mis à jour la taxonomie associée à chacune des séquences qui constituent les 4 bases de données de référence que j'ai utilisées (voir ci-dessous), avec la taxonomie à jour[3] du NCBI. Dans certains cas, la base de données permettait d'obtenir directement l'identifiant taxonomique du NCBI, auquel cas il suffisait de récupérer la taxonomie complète à l'aide de la fonction `lineage` de l'outil `taxonkit`. Dans d'autres cas, il fallait utiliser le nom de l'espèce pour retrouver l'identifiant taxonomique, soit par la fonction `name2taxid` de `taxonkit`, soit en cas de changement du nom de l'espèce, en utilisant les requêtes en ligne via la suite `Entrez` et son interface sur la librairie `biopython`.

Voici les 4 bases de données de référence utilisées.

Pour les données 16S : j'ai utilisé la base de données SILVA [142], dans sa version 138.1, en n'utilisant que les séquences avec un score de qualité (*pintail*) égal à 100, contenant $\sim 317K$ séquences.

Pour les données de cpn60 : j'ai utilisé la base de données cpnDB [68, 179] en utilisant les séquences du groupe I (bactéries et archées) avec toutes les séquences correspondant

3. en date du 25 mars 2022

aux cibles universelles (celles que nous utiliserons), contenant $\sim 17K$ séquences. Les identifiants taxonomiques n'étant pas disponibles dans les fichiers téléchargeables, j'ai sollicité Pr. Janet Hill, curatrice de la base de données, qui m'a transmis cette information lorsqu'elle était disponible. J'ai complété l'information à l'aide de la méthodologie présentée précédemment, et je l'ai transmise, à sa demande, au Pr. Janet Hill, contribuant ainsi à la curation de la base de données en fournissant l'identifiant taxonomique des 1127 séquences qui manquaient (4% de la base de données).

Pour les données de gyrB : j'ai utilisé la base de données construite par les auteurs d'une publication qui évalue ce marqueur et le compare au 16S [137], dont certains travaillent au sein de l'unité MaIAGE. Je remercie notamment Olivier Rué pour son assistance. Cette base de données contient $\sim 97K$ séquences.

Pour les données d'ITS$_{bact}$: j'ai utilisé la base de données construite par les auteurs qui ont développé cette approche [111, 110]. Cette base de données contient $\sim 129K$ séquences.

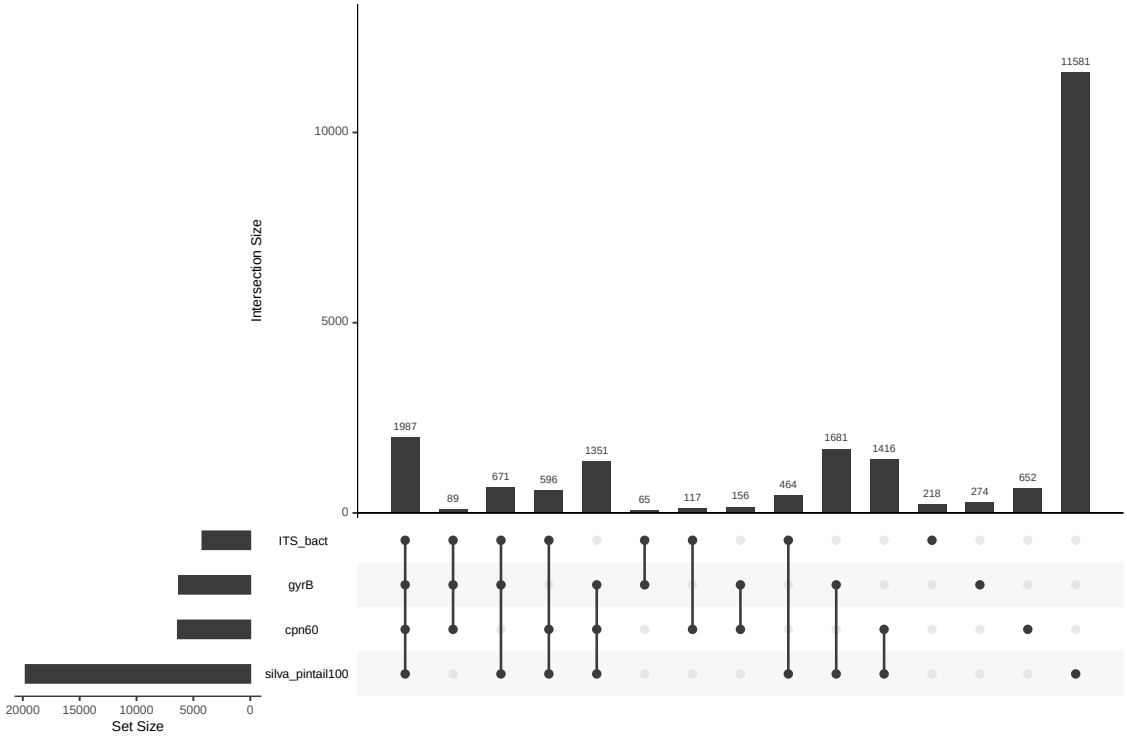


FIGURE 3.10 – Comparaison de la composition en espèces des bases de données de référence utilisées pour réaliser les assignations taxonomiques, pour les différents marqueurs. Dans ce graphique, de type *upset plot*, le nombre d'espèces de chaque base de données est représenté dans l'histogramme de gauche, et le nombre d'espèces retrouvées exclusivement dans chaque intersection de bases de données apparaît dans l'histogramme du haut. Par exemple, 218 séquences sont présentes uniquement dans ITS$_{bact}$ tandis que 1987 sont présentes dans toutes les bases.

On observe sur la figure 3.10 que SILVA, la base de données de référence que j'ai utilisée pour le gène 16S, est de très loin la plus complète. Elle contient beaucoup plus d'espèces bactériennes ($\sim 20K$) que les autres, et c'est un des arguments qui fait la popularité du 16S. Les autres

bases de données contiennent $\sim 5K$ espèces bactériennes différentes chacune, dont $\sim 3K$ qui sont partagées entre toutes les bases de données. Les espèces cliniquement importantes ne sont pas toutes représentées dans les bases de données. J'ai ensuite observé que parmi les espèces d'intérêt clinique du microbiote vaginal listées dans la table 3.2, BVAB1 était la seule espèce qui était absente des 4 bases de données, ce qui s'explique par le fait qu'elle n'ait jamais été cultivée donc difficilement répertoriée. SILVA contient toutes les autres espèces, alors que les bases de données de référence pour gyrB, cpn60 et ITS_{bact} ne contiennent pas respectivement une (*S. sanguinegens*), deux (*M. hominis* et *U. urealyticum*) et trois (*S. vaginalis*, *S. sanguinegens* et *M. indolicus*) autres espèces.

Pipeline bioinformatique Le pipeline bioinformatique que j'ai utilisé pour analyser chaque marqueur, présenté dans la figure 3.11, est le suivant. Les amorces sont retirées des *reads*, puis un filtre qualité est appliqué pour retirer les *reads* de mauvaise qualité[4]. Un modèle d'erreur est ensuite ajusté aux données afin de débruiter les séquences qui sont susceptibles de résulter du bruit introduit par les erreurs de séquençage et de se ramener à un ensemble de séquences uniques (Amplicon Sequence Variant, ou ASV). Les *reads* pour lesquels il y a un chevauchement suffisant (12 nucléotides, paramètre par défaut de DADA2) sont assemblés et les autres simplement concaténés avec une séquence de 10 "N" entre les deux. Les séquences chimériques sont ensuite recherchées et retirées. Les séquences des ASVs pour lesquelles les *reads* ont pu être assemblés sont alignés contre la base de données de référence avec `blastn`. Concernant les ASVs pour lesquelles les *reads* ne sont pas assemblées, les *reads* R1 et R2 sont alignés indépendamment sur la base de données, avec les mêmes paramètres, puis les résultats de ces alignements sont croisés pour obtenir des statistiques "globales" de l'alignement. Pour chaque séquence de la base de données de référence identifiées à la fois par l'alignement de R1 et de R2, on définit le nombre de mismatchs total de l'alignement comme la somme des mismatchs de l'alignement de R1 et de R2. La longueur totale de l'alignement est calculée de manière similaire alors que le pourcentage d'identité est lui défini comme la moyenne du pourcentage d'identité de l'alignement de R1 et de R2.

J'ai ensuite développé un script `R` qui permet de déterminer une assignation taxonomique pour chaque ASV à partir des résultats de l'alignement. Pour chaque ASV, j'ai récupéré les *hits* vérifiant (i) $\ell \geq 0.9 * \ell_{max}$ avec ℓ la longueur d'alignement pour chaque *hit* et ℓ_{max} la longueur d'alignement maximale trouvée dans l'alignement, et (ii) $m \leq m_{min} + 1$ avec m le nombre de *mismatchs* entre la séquence de l'ASV et le *hit* et m_{min} le nombre de *mismatchs* du meilleur *hit*. À chaque niveau taxonomique, j'ai assigné l'ASV à un taxon si tous les meilleurs *hits* appartenaient à ce taxon. Dans le cas contraire, l'ASV est noté comme "multi-affilié". Dans les cas où certains taxa n'étaient pas informatifs ("*unknown*", "unclassified", *etc.*), je ne les ai pas pris en compte pour établir l'assignation.

Le pipeline est largement inspiré de FROGS [44, 14], un outil développé au sein d'INRAE dédié à l'analyse de données de métabarcoding. Il avait dans un premier temps été décidé d'utiliser FROGS, mais il est apparu essentiel d'utiliser DADA2 pour le débruitage [25], outil qui n'est pas disponible dans FROGS à l'heure actuelle. En effet, l'utilisation d'ASVs (issus de DADA2), plutôt que d'OTUs (issus de l'outil Swarm [99] dans le pipeline FROGS) est préférable dans notre cas, à cause de la grande similarité entre les espèces du genre *Lactobacillus* (cf section 3.1.4). Il était alors plus simple de réaliser l'ensemble de l'analyse (pré-traitement, assemblage et suppression des chimères) avec DADA2, même si FROGS permet de réaliser les mêmes

4. `dada2::filterAndTrim(... , maxN=0, maxEE=2, truncQ=2, rm.phix=TRUE)`

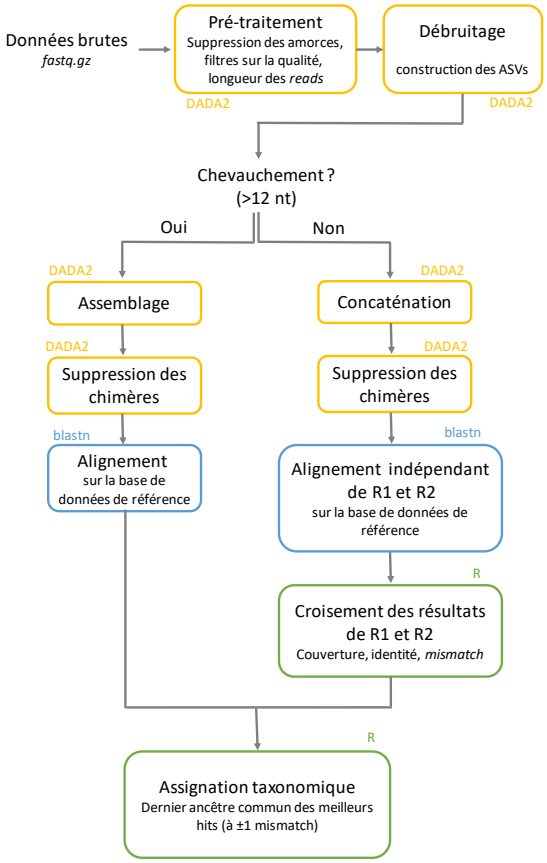


FIGURE 3.11 – Pipeline d'analyse de données métabarcoding pour chaque marqueur.

étapes. La gestion des *reads* qui ne sont pas assemblés, l'outil utilisé pour l'alignement, ainsi que les paramètres sont ceux utilisés par FROGS. Le script que j'ai utilisé pour l'assignation taxonomique permet de réduire le nombre de "multi-affiliation", sans passer par une curation manuelle des résultats de l'alignement, comme proposée dans FROGS.

3.3.2 . Méthodes pour obtenir le profil taxonomique consensus

À partir de maintenant, nous supposons que nous disposons d'autant de profils taxonomiques de notre écosystème que de marqueurs étudiés. Il s'agit donc dans ce paragraphe de regarder comment combiner ces différents profils pour en construire un unique, appelé profil taxonomique consensus.

On note par la suite r_i^m le comptage (nombre de *reads*) du taxon i pour le marqueur m. Lorsque deux taxa sont descendants immédiats d'un même noeud, on les désignera par i et j tandis que lorsqu'un taxon est parent immédiat de l'autre, on les désignera respectivement par p et e. L'ensemble des enfants immédiats d'un taxon p est noté $\mathcal{E}(p)$.

Processus récursif Pour obtenir le profil taxonomique consensus, j'ai développé une approche récursive qui résout la composition du rang taxonomique du plus élevé au plus bas (du domaine, phylum, vers l'espèce voire la sous-espèce). Le principe est le suivant.

Lorsque l'on cherche à résoudre la composition à un rang taxonomique donné (par exemple l'espèce), on suppose que l'on a préalablement résolu l'abondance relative consensus des taxa

du rang taxonomique supérieur, alors dénommés parents (dans cet exemple, les genres). Pour chaque parent, se pose alors un problème indépendant pour déterminer l'abondance relative consensus de ses enfants (toutes les espèces appartenant à un genre donné).

L'initialisation du processus se fait en définissant que l'abondance relative de la racine de l'arbre taxonomique vaut 1, puis en cherchant la composition en domaine.

Pour résoudre chaque itération, j'ai proposé 3 méthodes détaillées ci-dessous (voir figure 3.12) et qui seront évaluées dans les sections 3.4.1 et 3.4.2. Les méthodes C1 et C3 sont basées sur des transformations classiquement utilisées en analyse de données compositionnelles (*alr* en utilisant successivement tous les taxa comme références pour C1, *clr* pour C3) [2, 61, 60]. La méthode C2 est basée sur l'observation empirique que les enfants non-amplifiées par certains marqueurs, ou non assignés au niveau taxonomique des enfants, ont des abondances relatives biaisées vers le bas. Ainsi, prendre la valeur maximale du ratio enfant/parent à travers les différents marqueurs, plutôt que la moyenne, pourrait être intéressante.

Méthode C1 : ratios entre enfants, méthode *alr* avec références successives La première méthode compare les enfants d'un même parent p entre eux : pour chaque paire d'enfants $(i, j) \in \mathcal{E}(p)^2$, on calcule une moyenne (pondérée ou non), sur chaque marqueur m, du log ratio entre les comptages r_i^m et r_j^m des enfants i et j :

$$\rho_{ij} = \frac{1}{\sum_m \omega_{ij}^m} \sum_m \omega_{ij}^m log_2 \frac{r_j^m}{r_i^m}$$

avec ω_{ij}^m la pondération associée. Dans un premier temps, cette pondération vaut 1 si $r_i^m r_j^m > 0$ et 0 sinon, autrement dit on ne considère que les marqueurs m ayant identifié les deux enfants. Voir la figure 3.12 pour un exemple numérique.

La matrice carrée $(\rho_{ij})_{ij}$ obtenue contient ainsi les log ratios moyennés sur les marqueurs entre chaque paire d'enfants avec $\rho_{ii} = 1$ et ρ_{ji} = -ρ_{ij}. En particulier, la i^{eme} ligne $(\rho_{ij})_{j \in \mathcal{E}(p)}$ correspond au profil *alr* moyen obtenu en utilisant le taxon i comme référence. Pour des soucis de lisibilité et d'interprétation, cette matrice est représentée sous forme de ratio et non de log ratio dans la figure 3.12. Nous cherchons alors un vecteur d'abondance relative des enfants qui respecte au mieux cette matrice. La ligne i de cette matrice peut être vue comme l'abondance relative, à une constante multiplicative près, qu'on obtiendrait si on considérait l'enfant i comme référence, le définissant à 1. Ainsi, après passage à l'exponentielle, on peut diviser chaque ligne de la matrice (ρ_{ij}) par sa somme, pour obtenir une estimation du vecteur d'abondance relative recherché :

$$\gamma_{ij} = \frac{2^{\rho_{ij}}}{\sum_j 2^{\rho_{ij}}}.$$

Chaque ligne de la nouvelle matrice (γ_{ij}) peut alors être vue comme une estimation du vecteur de l'abondance relative des enfants, en considérant une référence différente par ligne, et nous en faisons une moyenne :

$$\tau_j = \frac{1}{|\mathcal{E}(p)|} \sum_i \gamma_{ij}.$$

Méthode C2 : ratios enfant/parent La deuxième méthode que nous avons proposée consiste à calculer les ratios enfant/parent r_e^m / r_p^m pour tous les enfants e du parent p, et à garder pour chaque enfant la valeur maximale de ce ratio parmi les différents marqueurs :

$$\rho_e = max_m \log_2 \left(\frac{r_e^m}{r_p^m} \right).$$

Il s'agit là encore d'une agrégation des profils *alr*, comme dans la méthode C1 mais avec deux différences majeures : (i) la référence choisie est le parent et (ii) l'agrégation est faite en prenant le maximum coordonnée par coordonnée (plutôt que la moyenne).

On normalise enfin ce vecteur ρ en le divisant par la somme de ses composantes pour obtenir le vecteur d'abondance relative ; voir la figure 3.12.

Remarque : le nombre de *reads* r_p^m assignés au parent p par le marqueur m est égal à la somme des r_e^m pour tous les enfants e du parent p additionnée du nombre de *reads* assignés jusqu'au rang du taxon p mais non assignés à un rang plus précis. En particulier, $r_p^m \geq \sum_{e \in \mathcal{E}(p)} r_e^m$ avec égalité si et seulement tous les *reads* assignés à p le sont également à un de ses enfants e.

Méthode C3 : transformation *clr* Enfin, nous proposons une troisième méthode, basée cette fois ci sur la transformation *centered log ratio* (*clr*). Les données sont tout d'abord transformées en profils *clr*, qui ne nécessitent pas de référence, avant d'être moyennées avec ou sans pondération :

$$x_i = \frac{1}{\sum_m \omega_m} \sum_m \omega_m log \frac{r_i^m}{\mu_m} \quad \text{avec} \quad \mu_m = \left(\prod_{i : r_i^m \neq 0} r_i^m \right)^{1/\#\{i : r_i^m \neq 0\}}$$

où μ_m est la moyenne géométrique du nombre de *reads* assignés à chaque enfant [5] pour le marqueur m et $(\omega_m)_m$ un ensemble de poids.

Enfin, on obtient l'abondance relative des enfants en réalisant l'opération inverse de la *clr* :

$$\tau_i = \frac{e^{x_i}}{\sum_{i'} e^{x_i'}}$$

qui permet de construire un vecteur d'abondance relative dont les composantes somment à 1. (voir la figure 3.12)

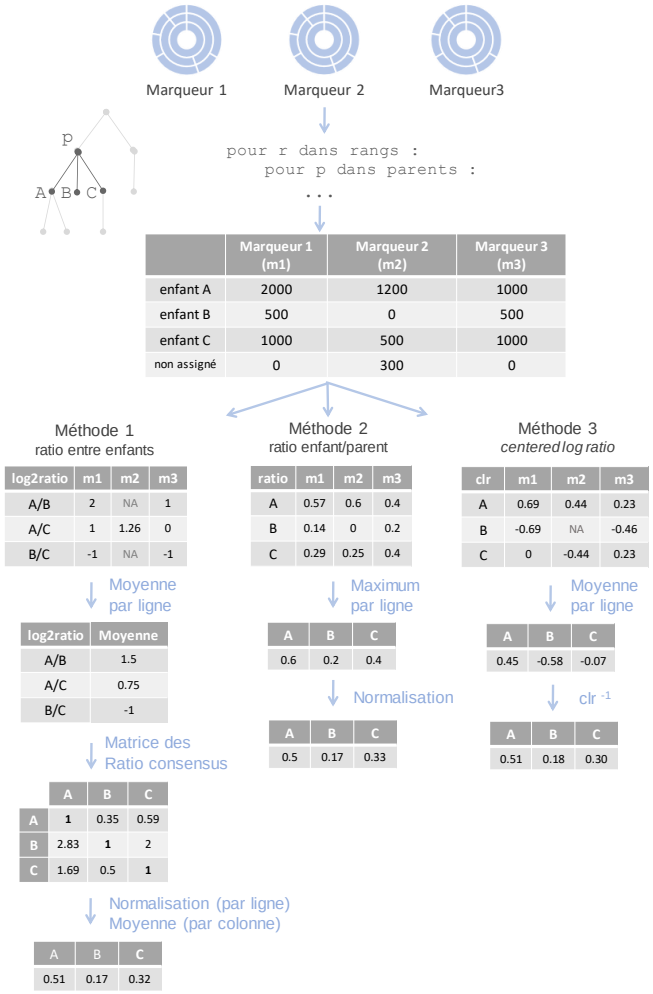
Pondération Les méthodes 1 et 3, qui utilisent une moyenne à travers les valeurs issues des différents marqueurs, peuvent faire intervenir une pondération de ces marqueurs. Dans certains cas, les données nous permettent d'évaluer la pertinence de chaque marqueur pour évaluer les abondances relatives à l'intérieur d'une clade. Le système de pondération que j'ai testé vise à évaluer le pouvoir discriminant de chaque marqueur pour la clade et ainsi à donner plus de poids aux marqueurs qui sont les plus résolutifs. J'ai donc défini la pondération comme le taux de *reads* assignés aux enfants :

$$\omega_m = \frac{\sum_i r_i^m}{r_p^m}.$$

3.3.3 . Construction des jeux de données simulés

Pour développer la méthode de construction du profil taxonomique consensus, j'ai d'abord constitué un jeu de données simulé simplifié me permettant de tester finement l'influence des différents biais énoncés plus haut, puis j'ai simulé un jeu de données plus réaliste de l'écosystème vaginal me permettant d'évaluer la méthodologie en attendant l'arrivée des données en cours de production à Alphabio. Ces deux jeux de données simulés sont décrits ci-dessous.

5. Nous n'utilisons pas de *pseudo-comptage*, les comptages nuls sont retirés et ne contribueront pas au calcul du consensus.



	Marqueur 1 (m1)	Marqueur 2 (m2)	Marqueur 3 (m3)
enfant A	2000	1200	1000
enfant B	500	0	500
enfant C	1000	500	1000
non assigné	0	300	0

log2ratio	m1	m2	m3
A/B	2	NA	1
A/C	1	1.26	0
B/C	-1	NA	-1

log2ratio	Moyenne
A/B	1.5
A/C	0.75
B/C	-1

	A	B	C
A	1	0.35	0.59
B	2.83	1	2
C	1.69	0.5	1

A	B	C
0.51	0.17	0.32

ratio	m1	m2	m3
A	0.57	0.6	0.4
B	0.14	0	0.2
C	0.29	0.25	0.4

A	B	C
0.6	0.2	0.4

A	B	C
0.5	0.17	0.33

clr	m1	m2	m3
A	0.69	0.44	0.23
B	-0.69	NA	-0.46
C	0	-0.44	0.23

A	B	C
0.45	-0.58	-0.07

A	B	C
0.51	0.18	0.30

FIGURE 3.12 – Représentation schématique des 3 méthodes qui permettent d'obtenir le profil taxonomique consensus (les valeurs numériques sont un exemple fictif). Le terme *normalisation* désigne ici la division d'un vecteur par sa somme.

Jeu de données simpliste Ce premier jeu de données est constitué de 6 profils types, dont les compositions sont détaillées dans la table 3.5. Si le premier profil considère les 4 espèces de Lactobacillus comme équidistribuées, les 5 profils suivants font directement référence aux 5 classes de microbiote vaginal vues au paragraphe 3.1.4 page 57.

Pour chaque profil type, j'ai constitué des échantillons avec plusieurs niveaux de difficulté, présentés dans la table 3.6, selon qu'ils présentent un biais du nombre de copies et/ou un biais d'universalité. Dans le niveau de difficulté le plus simple, noté A, seule la différence de pouvoir discriminant est à l'origine de la différence entre les profils issus des différents marqueurs, alors que les biais d'universalité des amorces et du nombre de copies présents dans les niveaux de difficultés de B à D vont rajouter des discordances entre les profils et complexifier la tâche d'intégration en un profil consensus.

Pour chacun des 5 marqueurs, chaque profil type et chaque niveau de difficulté, j'ai simulé le séquençage avec $10K$ *reads* appariés de $2*251$ bp, avec `grinder` [9]. Pour ces simulations, j'ai utilisé les même génomes que ceux utilisés dans les travaux préliminaires (cf section 3.2, page 61). J'ai introduit un bruit dans les données de séquençage avec un modèle d'erreur empirique

Nom	Description	*L. crispatus*	*L. gasseri*	*L. iners*	*L. jensenii*	Autres espèces d'intérêt
s0	Mélange de *Lactobacillus* spp.	25%	25%	25%	25%	
s1	$\sim$ CST I	90%				10%
s2	$\sim$ CST II		90%			10%
s3	$\sim$ CST III			90%		10%
s4	$\sim$ CST IV					100%
s5	$\sim$ CST V				90%	10%

TABLE 3.5 – Composition des profils types qui ont servi au développement de la méthode.

Niveau de difficulté	Biais du nombre de copies	Biais d'universalité
A	Non	Non
B	Non	Oui
C	Oui	Non
D	Oui	Oui

TABLE 3.6 – Niveaux de difficulté et biais associés.

basé sur des données obtenues à Alphabio en séquençage 16S. Le taux d'erreur croît le long du *read* et sa moyenne est de 4.10^{-4}.

Le biais du nombre de copies est facilement contrôlable par un paramètre de `grinder`. Pour générer des échantillons sans biais d'universalité des amorces (niveaux de difficulté A et C), j'ai retiré des compositions théoriques tous les génomes pour lesquels au moins un marqueur n'était pas amplifié. Ainsi, les échantillons des niveaux A et C sont restreints aux 4 espèces prépondérantes de *Lactobacillus* et à 2 autres espèces (*E. lenta* et *P. timonensis*). Il faut noter que `grinder` ne tolère aucun *mismatch* entre les amorces et les génomes pour l'amplification, contrairement à ce qui se passe en conditions réelles. On peut donc supposer que le biais d'universalité sera plus important dans les simulations qu'en conditions réelles.

Jeu de données réaliste J'ai récupéré les compositions de microbiotes vaginaux disponibles via le package R `curatedMetagenomicData` (N=96), incluant 67 échantillons du *Human Microbiome project* [172] réalisé aux Etats-Unis en 2012, 10 échantillons provenant d'une étude [184] menée au Luxembourg en 2018, et 19 échantillons issus d'une cohorte italienne [50] établie en 2018. Au sein de ces échantillons, 194 espèces différentes avaient été identifiées, parmi lesquelles je n'ai gardé que les 167 qui avaient un génome représentatif dans RefSeq. Basé sur ces compositions en espèces, j'ai réalisé des simulations avec la méthodologie décrite précédemment, en utilisant les 5 marqueurs et les 4 niveaux de difficulté. Les 5 marqueurs ne sont simultanément amplifiés que sur 48 espèces parmi les 167. Par conséquent les profils limités à ces 48 espèces (niveaux de biais A et C) étaient très éloignés des profils initiaux, et donc peu pertinents pour l'analyse. Le niveau de difficulté D, avec biais d'universalité et biais du nombre de copies, est donc le plus réaliste et le plus pertinent pour évaluer la méthode ; les résultats montrés dans la section suivante correspondront au niveau de difficulté D, sauf mention explicite du contraire.

3.3.4 . Classification en CSTs

Comme mentionné en introduction de ce chapitre (cf. section 3.1.4, page 57), il y a 5 classes principales dans le microbiote vaginal, appelées CSTs. Le fait de pouvoir classer les échantillons selon ces 5 classes est important pour le diagnostique clinique. Parmi ces classes, la CST IV est à la fois la plus intéressante, car liée à des vaginoses, et la plus difficile à identifier car elle ne comporte pas d'espèce dominante de *Lactobacillus*. En effet, lorsqu'un échantillon contient une ou plusieurs espèces du genre *Lactobacillus*, mais pas en dominance nette par rapport aux autres espèces, cet échantillon peut être classé en CST IV ou dans une autre CST. Il n'y a pas de règle universelle pour déterminer l'appartenance d'un échantillon à une classe. Les méthodes décrites dans la littérature [144, 23] se basent sur une analyse non-supervisée d'un jeu de données entier (et non d'un échantillon isolé), pour déterminer des clusters, lesquels sont ensuite étiquetés selon les classes.
Pour déterminer les CSTs, j'ai réalisé une classification ascendante hiérarchique des échantillons, en utilisant la distance de Bray-Curtis et la méthode de Ward. Les distances de Bray-Curtis sont calculées

- à partir de la table de comptage des ASVs pour les approches mono-marqueur,
- à partir des résultats des méthodes de reconstruction du profil taxonomique consensus que j'ai développées, pour les approches multi-marqueurs,
- sur la composition théorique en espèces pour la classification de référence.

J'ai ensuite récupéré les principaux clusters afin d'identifier manuellement à quelle CST ils correspondaient, en regardant l'assignation taxonomique des espèces dominantes. Les clusters dominés par le genre *Lactobacillus* sont attribués aux CST I, II, III ou V selon l'espèce qui est identifiée. Si l'assignation taxonomique n'est pas faite au niveau espèce, le cluster n'est pas attribué à une CST. Les clusters qui ne sont pas dominés par le genre *Lactobacillus* sont attribués à la CST IV.

3.4 . Résultats

Dans cette section, nous allons d'abord nous intéresser au calibrage de la méthode, basé sur le jeu de données simpliste (section 3.4.1), en étudiant les propriétés des méthodes de consensus sur des critères de présence/absence des espèces, puis de l'estimation de leurs abondances relatives. Nous évaluerons ensuite les méthodes sur les jeux de données simulés réalistes (section 3.4.2).

3.4.1 . Calibrage de la méthode

Identification des *Lactobacillus* L'identification des espèces de *Lactobacillus* est un enjeu majeur dans le contexte qui nous intéresse, à savoir le microbiote vaginal. D'après notre travail préliminaire (cf. section 3.2, figure 3.9, page 65), on sait que les différents marqueurs n'ont pas le même pouvoir résolutif pour ce groupe taxonomique. En prenant le jeu de données *s0* correspondant au mélange équidistribué des 4 espèces d'intérêt de *Lactobacillus*, on pourra évaluer chaque marqueur ainsi que la capacité des méthodes multi-marqueurs qui combinent les marqueurs à tirer parti du pouvoir discriminant de chaque marqueur. On observe sur la figure 3.13 qu'aucun marqueur, considéré individuellement, ne permet de classifier toutes les séquences : cpn60, gyrB et ITS_{bact} ont identifiés les 4 espèces, mais n'ont pu assigner certaines séquences. Pour ces 3 marqueurs, l'assignation taxonomique au niveau de l'espèce n'a pas été possible pour certains génomes ou certaines copies du marqueur. Les marqueurs 16S V3V4 et 16S V1V4

n'ont permis d'identifier qu'une partie des espèces d'intérêt, la résolution taxonomique n'étant pas assez bonne pour les autres. À l'inverse, en combinant les 5 marqueurs, on arrive à résoudre totalement le profil taxonomique au niveau des espèces et à identifier ainsi les 4 *Lactobacillus* d'intérêt et ce, quelle que soit la méthode utilisée pour reconstruire le profil consensus.

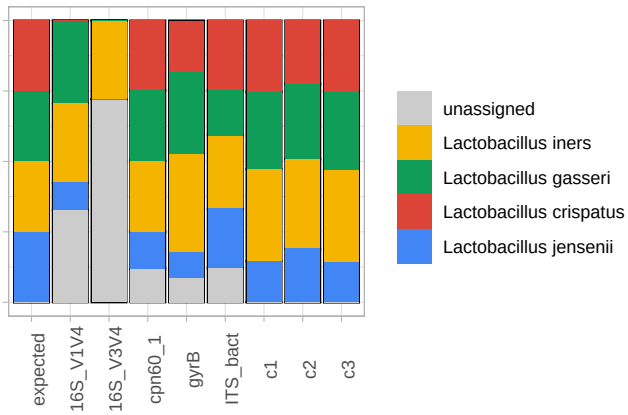


FIGURE 3.13 – Compositions en espèces attendues et prédites par les 5 marqueurs pris individuellement, puis par les différentes méthodes multi-marqueurs C1, C2 et C3 pour le jeu de données *s0* au niveau de difficulté D.

Identification des autres espèces du microbiote vaginal Intéressons-nous ici à la capacité de nos 5 marqueurs et des 3 méthodes multi-marqueurs à identifier les autres espèces d'intérêt du microbiote vaginal. Pour cela, la CST IV de microbiotes vaginaux, qui n'est pas dominée par le genre *Lactobacillus* et présente la plus grande biodiversité, est la plus pertinente. Nous considérons donc dans ce paragraphe le jeu de données *s4* au niveau de difficulté D. Dans cette situation, l'absence d'une espèce dans un profil peut s'expliquer par le manque de résolution taxonomique (l'assignation taxonomique n'est faite qu'à un niveau plus élevé), mais aussi par la non-amplification du marqueur. On évaluera ainsi la capacité des méthodes multi-marqueurs à tirer parti de la complémentarité des marqueurs dans ces conditions.

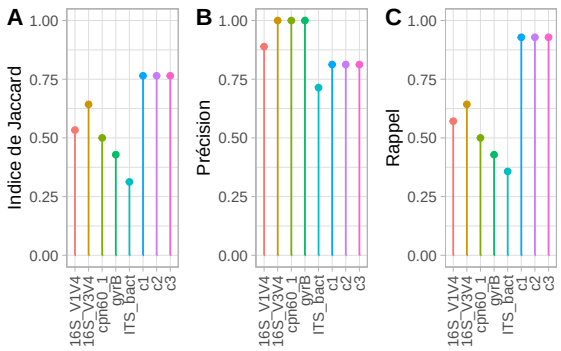


FIGURE 3.14 – Performance des différentes méthodes pour l'identification des espèces dans le jeu de données *s4* : (A) indice de Jaccard ($\frac{TP}{TP+FP+FN}$), (B) précision ($\frac{TP}{TP+FP}$) et (C) rappel ($\frac{TP}{TP+FN}$).

On observe sur la figure 3.14, que la composition en espèces obtenue est plus proche de la composition attendue, au sens de l'indice de Jaccard, en utilisant les méthodes multi-marqueurs que les méthodes à marqueur unique. La précision, qui pénalise les faux positifs, est moins élevée pour les méthodes multi-marqueurs que pour certains marqueurs pris individuellement. En effet, 16S V3V4, cpn60$_1$ et gyrB permettent d'obtenir des profils sans aucun faux positif, alors que les méthodes multi-marqueurs héritent des faux positifs identifiés par les marqueurs 16S V1V4 (*E. sp. mt102*) et ITS$_{bact}$ (*S. moniliformis* et *S. notomytis*). En termes de rappel, qui pénalise les faux négatifs, on observe que les méthodes multi-marqueurs font largement mieux que chaque marqueur pris individuellement. Les méthodes multi-marqueurs utilisent la complémentarité des marqueurs et parviennent ainsi à identifier correctement presque toutes les espèces d'intérêt. Seule *BVAB1*, espèce non encore cultivée et absente de toutes les bases de données utilisées, n'est pas identifiée par les méthodes mutli-marqueurs. On retrouve ici le compromis usuel : les gains en précision entraînent généralement une perte en rappel.

Estimation des abondances relatives des *Lactobacillus* dominants Nous allons désormais nous intéresser à l'estimation de l'abondance relative des *Lactobacillus* dans les jeux de données *s1*, *s2*, *s3* et *s5*. Dans ces jeux de données, on s'attend à retrouver 90% de *Lactobacillus*, comme mentionné dans la table 3.5. Il est intéressant de voir si les différentes méthodes permettent de bien estimer la dominance de *Lactobacillus*.

Les différents marqueurs donnent des résultats discordants sur l'abondance relative des taxa dès les rangs taxonomiques élevés (phylum), en raison du biais du nombre de copies, des différences de résolution taxonomique, et également du biais d'universalité des amorces. Par exemple, si un marqueur n'est pas amplifié chez certaines espèces d'un phylum, l'abondance relative de tout ce phylum sera sous-estimée.

	Difficulté	A				B				C				D			
	Jeu de données	*s1*	*s2*	*s3*	*s5*	*s1*	*s2*	*s3*	*s5*	*s1*	*s2*	*s3*	*s5*	*s1*	*s2*	*s3*	*s5*
mono-marqueur	16S V1V4	0.89	0.91	0.90	0.90	0.93	0.93	0.94	0.94	0.94	0.95	0.95	0.93	0.98	0.98	0.98	0.97
	16S V3V4	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.94	0.95	0.95	0.93	0.95	0.96	0.96	0.94
	cpn60$_1$	0.90	0.90	0.90	0.90	0.94	0.94	0.95	0.94	0.90	0.90	0.90	0.90	0.94	0.94	0.94	0.94
	gyrB	0.91	0.90	0.90	0.90	0.92	0.92	0.92	0.85	0.90	0.94	0.95	0.95	0.92	0.95	0.96	0.92
	ITS$_{bact}$	0.90	0.89	0.88	0.90	0.95	0.94	0.94	0.95	0.94	0.95	0.95	0.93	0.97	0.97	0.97	0.96
multi-marqueur	C1	0.90	0.90	0.89	0.90	0.93	0.93	0.93	0.92	0.93	0.94	0.94	0.92	0.96	0.97	0.96	0.95
	C1w	0.90	0.90	0.89	0.90	0.93	0.93	0.93	0.92	0.93	0.94	0.94	0.92	0.96	0.97	0.96	0.95
	C2	0.90	0.89	0.88	0.90	0.90	0.91	0.91	0.89	0.90	0.91	0.91	0.91	0.91	0.93	0.93	0.92
	C3	0.90	0.90	0.89	0.90	0.92	0.92	0.93	0.92	0.93	0.94	0.94	0.92	0.95	0.96	0.96	0.95
	C3w	0.90	0.90	0.89	0.90	0.92	0.92	0.93	0.92	0.93	0.94	0.94	0.92	0.95	0.96	0.96	0.95

TABLE 3.7 – Estimation de l'abondance relative des *Lactobacillus* par les différentes méthodes mono et multi-marqueurs pour les 4 jeux de données *s1*, *s2*, *s3* et *s5*, et pour les 4 niveaux de difficultés de 1 à D. La valeur attendue est 0.9, les valeurs en vert sont celles qui sont dans l'intervalle 0.9 ± 0.015, celles en jaune sont dans 0.9 ± 0.045 et les rouges hors de cette intervalle. Les méthodes C1w et C3w utilisent la pondération par le taux d'assignation taxonomique décrite dans la section 3.3.2

	attendue	16S V1V4	16S V3V4	$cpn60_1$	gyrB	ITS_{bact}	C1	C2	C3
Fusobacteria	1.4	0.6	0.4	0.0	0.6	0.5	0.6	0.6	0.7
Tenericutes	1.4	0.3	0.3	0.0	0.8	0.3	0.4	0.8	0.6
Actinobacteria	3.6	0.7	1.1	2.9	0.8	1.1	1.1	2.8	1.2
Bacteroidetes	1.4	0.3	0.3	1.5	0.0	0.2	0.4	1.5	0.4
Firmicutes	92.1	98.0	97.4	95.6	96.1	97.9	97.5	94.3	97.2
unassigned	.0	.0	0.5	0.0	0.8	0.0	0.0	0.0	0.0
Total	100.0	100.0	100.0	100.0	99.1	100.0	100.0	100.0	100.0

TABLE 3.8 – Abondances relatives des phyla, attendues et estimées par les différentes méthodes, dans le jeu de données *s2* au niveau de difficulté D.

On observe sur la table 3.7 que l'estimation de l'abondance relative des *Lactobacillus* est une tâche loin d'être évidente, et dépend fortement du niveau de biais.

Concentrons-nous dans un premier temps sur les approches monomarqueurs. Sans biais du nombre de copies ni d'universalité des amorces, *i.e.* niveau de difficulté A, on retrouve quasiment systématiquement les 90% attendus (voir table 3.5 pour les compositions attendues). Les écarts constatés s'expliquent par des séquences qui sont écartées lors du traitement bioinformatique : certaines séquences de ITS_{bact} provenant de *L. iners* sont identifiées à tort comme chimériques. Lorsque l'on introduit le biais d'universalité des amorces, niveau de difficulté B, on observe que seul 16S V3V4, étant amplifié chez tous les génomes d'intérêt, donne les résultats attendus. À l'inverse, lorsque l'on considère le seul biais du nombre de copies, niveau de difficulté C, c'est $cpn60_1$ qui donne les meilleurs résultats, étant donné qu'il est le seul marqueur présent en une seule copie par génome. En combinant ces deux sources de biais (niveau de difficulté D), ce qui est le cas le plus réaliste, on observe qu'aucun marqueur ne permet d'estimer correctement l'abondance relative des *Lactobacillus*.
Les méthodes multi-marqueurs estiment généralement mieux l'abondance relative des *Lactobacillus* que les approches mono-marqueurs. On observe notamment que la méthode C2 arrive à la bonne estimation dans l'ensemble des cas aux niveaux B et C, alors qu'aucun marqueur n'y parvient. Au niveau D, les méthodes multi-marqueurs ne permettent pas d'avoir une estimation parfaite, mais on note que la méthode C2 fait systématiquement mieux que tous les marqueurs pris individuellement.

Par ailleurs, on peut constater que la pondération n'a aucune influence sur les abondances relatives. J'ai également réalisé des essais avec les pondérations élevées au carré, sans que cela n'ait d'influence sur les résultats.

Cas problématiques pour l'estimation des abondances relatives Afin d'identifier les cas où les méthodes multi-marqueurs ne donnaient pas les résultats attendus et ainsi trouver des pistes d'amélioration, j'ai détaillé quelques cas difficiles. Le premier est l'évaluation de l'abondance relative des *Lactobacillus* dans les échantillons du jeu de données *s2* au niveau de difficulté D, qui est l'échantillon où les plus grands écarts à la valeur attendue ont été observés. J'ai constaté que ces écarts sont observés dès le niveau phylum. En effet, il y a une surestimation du phylum *Firmicutes*, auxquels appartiennent les *Lactobacillus*, mais le ratio entre *Lactobacillus* et *Firmicutes* est bon.
On observe sur la table 3.8 que chaque marqueur pris individuellement sur-estime les *Firmicutes*. Cette sur-estimation a des causes différentes. La composition donnée par 16S V3V4 souffre du biais du nombre de copies en faveur des *Firmicutes* (il donne en effet la valeur attendue aux

	expected	16S V1V4	16S V3V4	cpn60$_1$	gyrB	ITS$_{bact}$	C1	C2	C3
Sneathia	14.3	25.3	12	0	9.1	0.0	8.9	8.2	8.7
Streptobacillus	0.0	0.0	0	0	0.0	15.4	8.9	8.2	8.7
Total	14.3	25.3	12	0	9.1	15.4	17.7	16.4	17.5

TABLE 3.9 – Abondances relatives des genres au sein de la famille *Leptotrichiaceae*, attendues et estimées par les différentes méthodes, dans le jeu de données *s4* au niveau de difficulté D

niveaux A et B). cpn60 n'est amplifié que chez certaines bactéries des autres phyla, conduisant à sur-estimer les *Firmicutes* (il donne la valeur attendue aux niveaux A et C). Enfin, les autres marqueurs possèdent simultanément ces deux biais. Finalement, dans le scénario D, avec ce fonctionnement récursif en commençant par estimer la composition en phylum, il est difficile d'imaginer une méthode pour obtenir le consensus qui puisse trouver la bonne valeur de *Firmicutes*, sachant que chaque marqueur sur-estime cette valeur. Une solution serait de corriger les profils pour le biais du nombre de copies avant de réaliser l'intégration lorsque c'est possible.

Un autre exemple qui a montré les limites des méthodes multi-marqueurs est le jeu de données *s4*, représentant un mélange des espèces d'intérêt du microbiote vaginal, hors *Lactobacillus*, au niveau de difficulté D. Dans ce jeu de données, les abondances relatives obtenues ne sont pas exactement égales à celles attendues. Outre une surestimation des *Fusobacteria*, j'ai noté un problème au niveau de la famille *Leptotrichiaceae*.
On observe dans la table 3.9 qu'il y a une mauvaise classification dans le profil obtenu par ITS$_{bact}$. Après vérification, le genre *Sneathia* attendu est absent de la base données de référence pour ce marqueur ; les espèces associées *Sneathia vaginalis* et *Sneathia sanguinegens* sont en effet classifiées en tant que *Streptobacillus moniliformis* et *Streptobacillus notomytis*. Les méthodes qui cherchent le consensus donnent *Sneathia* et *Streptobacillus* en quantité égale, n'ayant aucun moyen d'évaluer l'abondance relative entre les deux genres car aucun marqueur ne détecte les deux. Cet exemple montre à quel point il est important d'avoir le moins possible de faux positifs dans les profils avant de faire la combinaison, quitte à être quelque peu conservateur dans l'assignation taxonomique. Dans cet exemple, les pourcentages d'identité entre les séquences observées et celles de la base de données sont relativement faibles, laissant penser qu'ajuster le script d'assignation taxonomique en n'autorisant l'assignation au genre que si le pourcentage d'identité dépasse un seuil permettrait d'améliorer le résultat. Cependant, il faudrait établir de tels seuils pour tous les marqueurs et niveaux taxonomiques, ce qui n'est pas simple. Une autre possibilité serait d'établir qu'un enfant n'est considéré comme présent que s'il est identifié par au moins deux, ou plus, marqueurs. Cela permettrait très certainement d'éliminer des faux positifs, mais pourrait écarter également des vrais positifs s'ils n'ont été identifiés que par un marqueur. Le bénéfice de cette solution dépendrait donc du taux de faux positifs dans les profils, mais également de l'appréciation de l'utilisateur et du choix des marqueurs, par exemple en sélectionnant au moins deux marqueurs universels ayant une base de données de référence suffisamment complète.

On peut voir dans ces exemples, qu'il n'y a pas (table 3.9) ou peu (table 3.8) de *reads* non assignés aux parents. Ceci explique pourquoi la pondération proposée, qui se base sur les taux d'assignation, ne permet pas d'améliorer les estimations des abondances relatives. Par conséquent, les méthodes avec pondération sont écartées de la suite de l'analyse.

3.4.2 . Évaluation de la méthode

Nous allons dans cette section évaluer la méthode en utilisant le jeu de données réaliste, simulé à partir des compositions de 96 échantillons de microbiotes vaginaux, comme expliqué dans le paragraphe 3.3.3, page 72.

Identification des taxa J'ai dans un premier temps agrégé les données des 96 échantillons et calculé les métriques globales de performances des méthodes, en terme de sensibilité, de rappel ainsi que l'indice de Jaccard, aux différents niveaux taxonomiques. Conformément à ce qui a été obtenu lors du premier lot de simulations (cf. section 3.4.1), on observe une amélioration importante pour l'identification des espèces présentes par les méthodes multi-marqueurs (voir figure 3.15). Ces dernières permettent de diminuer de manière importante les faux négatifs (meilleur rappel), sans dégrader massivement la précision (précision modérément plus faible), résultant ainsi en une composition plus proche de celle attendue (indice de Jaccard plus élevé), et ce quelque soit le niveau taxonomique. On observe par ailleurs que par rapport au premier jeu de données (cf. figure 3.14), il y a un taux de rappel moins important (entre 0.3 et 0.5 contre entre 0.35 et 0.65) pour les méthodes utilisant un seul marqueur, témoignant d'un plus grand taux de faux négatifs. Ceci est dû au fait que les marqueurs n'ont pas été amplifiés dans un grand nombre de génomes, comme précédemment discuté (cf. paragraphe 3.3.3, page 72). Ces indicateurs qualitatifs ne permettent pas de discriminer les méthodes multi-marqueurs entre elles, celles-ci ne différant que par leurs estimations des abondances relatives.

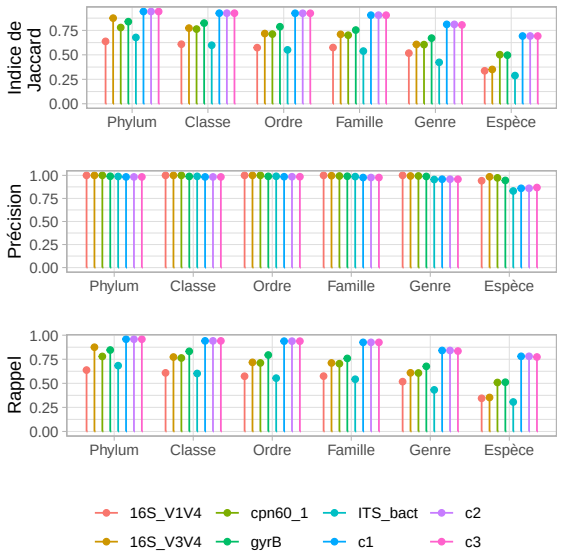


FIGURE 3.15 – Performance de l'identification des taxa sur le jeu de données simulé constitué de 96 échantillons du jeu de données simulé : indice de Jaccard ($\frac{TP}{TP+FP+FN}$), précision ($\frac{TP}{TP+FP}$) et rappel($\frac{TP}{TP+FN}$).

Estimation des abondances relatives L'évaluation quantitative des méthodes se fait en mesurant la distance entre les compositions théoriques et celles obtenues selon les différentes méthodes, en utilisant les distances de Jaccard (mesure de distance qualitative, pour confirmer les résultats précédents), de Bray-Curtis et d'Aitchison (mesures quantitatives, *cf.* section 1.3.1). Les distances mesurées par rapport au profil théorique sont très variables selon les échantillons,

ce qui rend la visualisation et l'interprétation compliquées. J'ai regardé dans un premier temps la distance moyenne, au niveau des espèces, sur tous les échantillons et classé les méthodes selon cette moyenne, comme montré dans la figure 3.16.

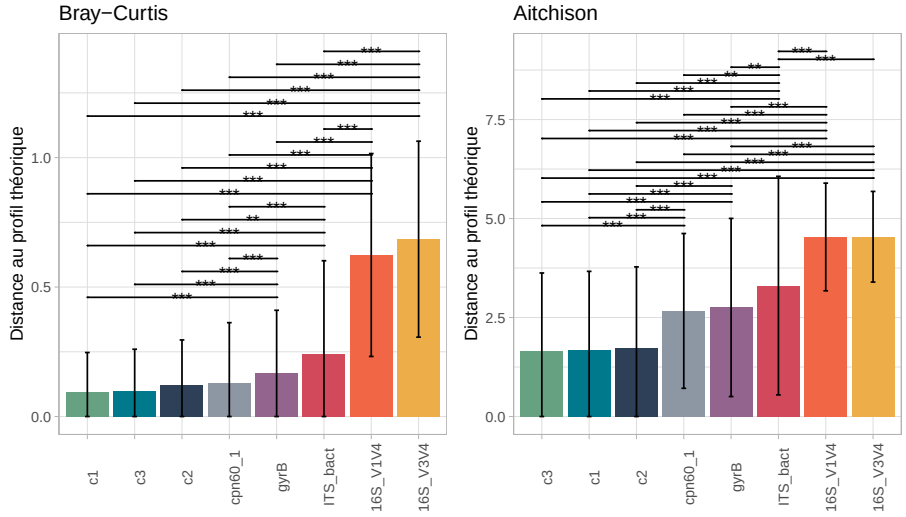


FIGURE 3.16 – Distances de Bray-Curtis et d'Aitchison entre les compositions en espèces attendues et celles mesurées par les différentes méthodes : moyenne ± écart-type sur les 96 échantillons du jeu de données simulé. Les méthodes sont ordonnées selon leur performance (les plus performantes sont celles qui minimisent la distance). Les barres horizontales, ainsi que les étoiles, représentent les résultats de comparaison entre chaque couple de méthodes (les extrémités indiquent les 2 méthodes comparées) par un test de Wilcoxon, apparié sur les échantillons. Seules les comparaisons dont les résultats sont significatifs apparaissent, avec les codes suivants : (*), (**) et (***) signifient que la p-value est inférieure à 0.05, 0.01, et 0.001 respectivement.

On observe sur la figure 3.16 que les profils obtenus par les méthodes multi-marqueurs sont plus proches en moyenne que ceux obtenus par les approches avec un seul marqueur, selon les distances de Bray-Curtis et d'Aitchison calculées sur les profils taxonomiques au niveau des espèces. Parmi les approches avec un seul marqueur, c'est cpn60 qui produit les meilleurs résultats. La différence de performance entre chaque paire de méthodes à été évaluée par un test de Wilcoxon apparié [6]. On observe qu'il n'y a pas de différence statistiquement significative entre les méthodes multi-marqueurs, qui font chacune mieux que toutes les méthodes basées sur un seul marqueur, à l'exception de cpn60, pour lequel la différence n'est pas significative selon la distance de Bray-Curtis. On note que les marqueurs 16S V1V4 et V3V4 sont très peu performants selon ces métriques, ceci étant dû au fait qu'ils ne permettent pas souvent une assignation taxonomique jusqu'à l'espèce.

J'ai ensuite regardé, pour chaque échantillon, quelle méthode donnait le meilleur résultat, c'est-à-dire la méthode qui minimise la distance entre le profil attendu et celui mesuré, selon les différentes métriques et niveaux taxonomiques, et j'ai décomposé le résultat selon les CSTs d'appartenance des échantillons, comme montré sur la figure 3.17. La classification en CSTs a été ici obtenue à partir des compositions théoriques (*cf.* section 3.3.4). On note dans un pre-

6. On a pour chaque méthode, et pour chaque échantillon la mesure de distance entre le profil observé et le profil attendu, et on compare les deux méthodes en appariant sur les échantillons

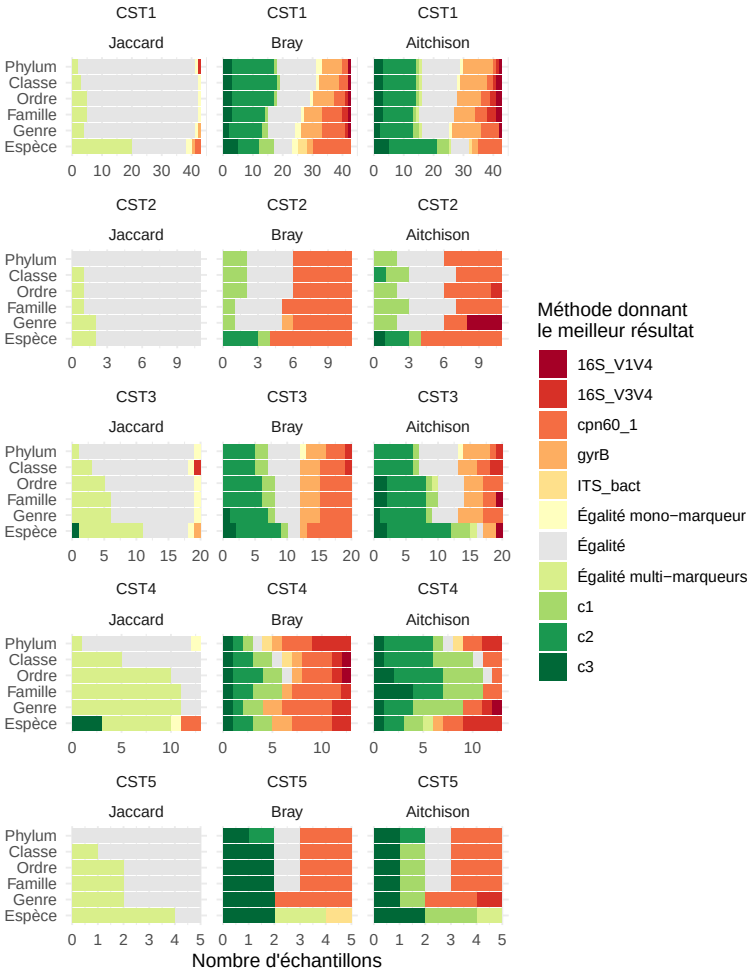


FIGURE 3.17 – Méthodes ayant produit le profil le plus proche du profil attendu, selon les distances de Jaccard, de Bray-Curtis et d'Aitchison, en fonction des rangs taxonomiques et des CSTs. En cas d'égalité, les situations où les meilleures méthodes sont exclusivement multi-marqueurs ou mono-marqueur sont distinguées dans des catégories à part.

mier temps que, sur un critère présence/absence des taxa (distance de Jaccard), les méthodes multi-marqueurs font au moins aussi bien (égalité) ou mieux que les méthodes mono-marqueurs. Lorsque qu'on s'intéresse aux abondances relatives des taxa (distances de Bray-Curtis et d'Aitchison), les résultats sont plus mitigés : il y a environ autant d'échantillons pour lesquels le meilleur résultat est obtenu par une méthode multi-marqueurs que par une approche basée sur un seul marqueur. On note que c'est souvent cpn60 qui produit les meilleurs résultats pour les CST II, III, V et en moindre mesure pour la CST I, probablement car ces CST sont dominées par une espèce du genre *Lactobacillus*, et que cpn60 discrimine correctement ces espèces et n'a pas de biais du nombre de copies, conformément à ce qui a été observé dans la section 3.4.1, table 3.7.

Classification en CST La classification en CSTs est une étape primordiale pour l'analyse du microbiote vaginal. Il est intéressant de comparer, selon les méthodes utilisées pour obtenir les profils taxonomiques, les résultats de cette classification. Nous avons donc réalisé la classification du jeu de données en CSTs, comme décrit dans la section 3.3.4.

	CST1 (N=43)	CST2 (N=11)	CST3 (N=20)	CST4 (N=13)	CST5 (N=5)
16S V1V4	0	0.64	1	0.85	0.0
16S V3V4	0	0	1	0.69	0.0
$cpn60_1$	1	0.91	1	1	0.0
gyrB	1	0.91	1	1	0.0
ITS_{bact}	1	0	1	0.54	1.0
C1	1	1	1	0.92	0.8
C2	1	1	1	0.85	1.0
C3	1	1	1	0.92	0.8

TABLE 3.10 – Taux de rappel ($\frac{TP}{TP+FN}$) de la classification en CST obtenue selon les différentes méthodes. La classification de référence qui sert à évaluer les autres est celle obtenue à partir des compositions théoriques des échantillons.

Notant que la CST I était souvent divisée en deux clusters, j'ai extrait les 6 clusters principaux, et non les 5 clusters principaux correspondants au 5 CSTs. En effet, au sein de la CST1, on retrouve un cluster avec uniquement des *L. crispatus* et un autre avec une dominance de *L. crispatus* mais une présence non-négligeable de *L. iners*.

Dans un premier temps, on observe sur la figure 3.18 que la structure en 5 clusters (ou 6 avec la sous-division de la CST I) est très clairement retrouvée par toutes les méthodes. Cependant, dû au fait que l'assignation taxonomique n'est pas toujours possible au niveau des espèces, certains clusters ne sont pas attribués à une CST. On observe qu'aucune des approches mono-marqueur ne permet l'identification de toutes les CSTs, à l'inverse des méthodes multi-marqueurs qui y parviennent.

En comparant la classification obtenue pour chaque échantillon par les différentes méthodes, table 3.10, on note que certains échantillons ne sont pas classés dans le même cluster que ce qui est obtenu avec les compositions théoriques. Les échantillons en question ont des *Lactobacillus* en abondance relative intermédiaire, et sont selon les méthodes retrouvées dans le cluster correspondant à la CST IV, ou dans un autre cluster.

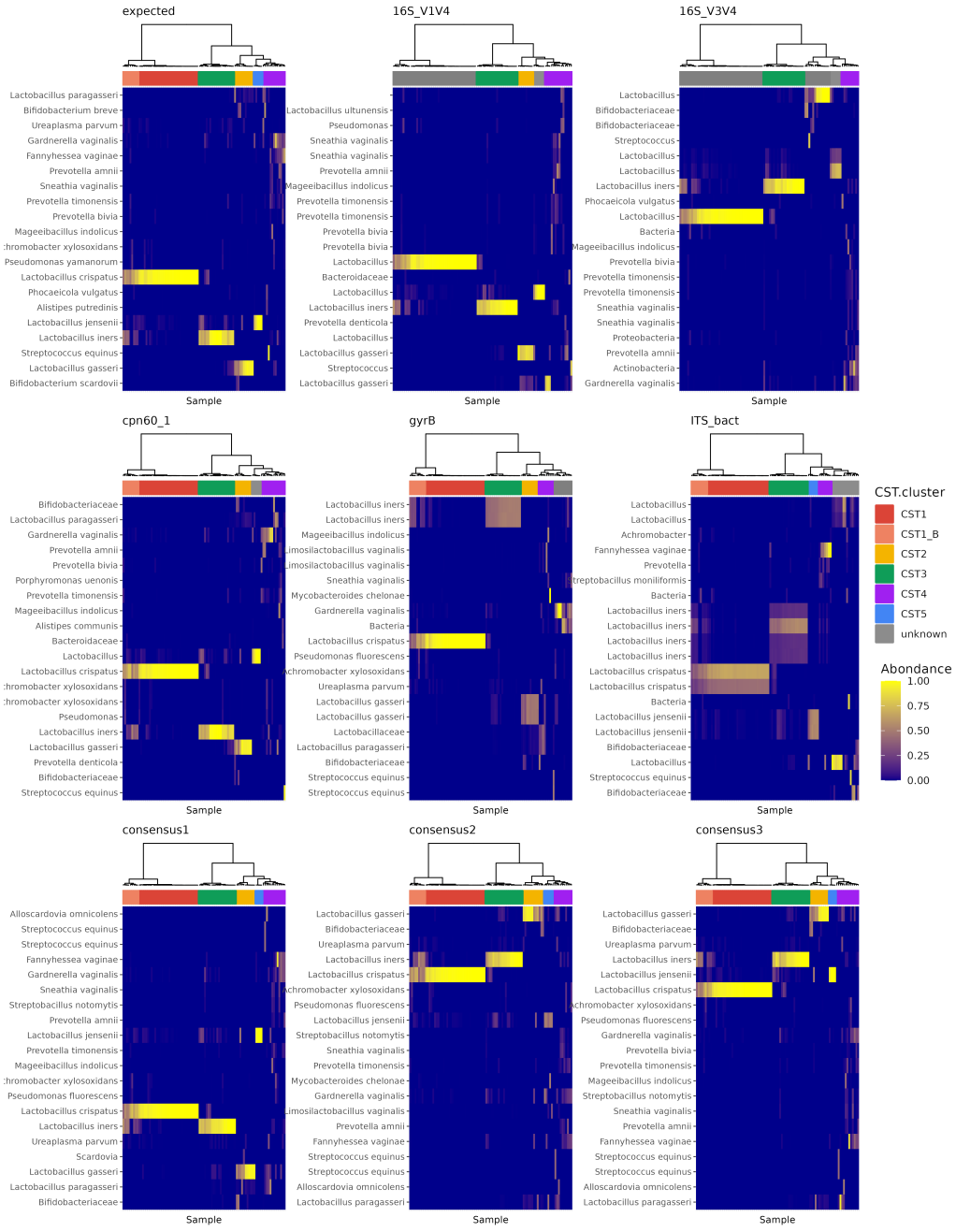


FIGURE 3.18 – Abondances relatives des entités taxonomiques obtenues pour les 96 échantillons du jeu de données simulé. Dans chaque vignette, le dendogramme du clustering hiérarchique est montré en haut, et la ligne supérieure représente la CST attribuée. Les lignes correspondent aux espèces dans les profils théoriques ; aux ASVs pour les approches mono-marqueur ; et aux niveaux taxonomiques les plus bas identifiés dans le profil taxonomique consensus pour les approches multi-marqueurs.

3.5 . Discussion

3.5.1 . Bilan

Pour caractériser le microbiote vaginal par métabarcoding, on a vu dans un premier temps qu'aucun gène marqueur ne permet à lui seul d'atteindre des résultats satisfaisants, mais que les différents marqueurs étaient complémentaires. Le gène 16S, grâce à ses régions conservées, permet la création d'amorces universelles et ainsi une évaluation globale de la composition de l'écosystème. Cependant, l'assignation taxonomique est souvent limitée au niveau du genre, notamment au sein des *Lactobacillus*, population clé dans le microbiote vaginal. Les autres marqueurs testés sont moins universels mais souvent plus résolutifs, notamment chez les *Lactobacillus*, et ont pour certains un biais du nombre de copies moins important, mettant ainsi en évidence leur complémentarité par rapport au 16S et donc l'intérêt de les combiner. On a également pu constater que les profils taxonomiques issus des différents marqueurs présentent certaines différences, et ce dès les niveaux taxonomiques élevés comme le phylum, du fait des biais du nombre de copies et d'universalité. L'approche que j'ai développée, consistant à séquencer plusieurs marqueurs, obtenir un profil taxonomique pour chaque marqueur, puis combiner ces profils pour obtenir un profil taxonomique consensus, permet une grande liberté de choix à l'utilisateur, tant sur le choix des gènes marqueurs que sur les outils et bases de données utilisés pour traiter les données de chaque marqueur. Pour construire les profils consensus, étape qui constitue le coeur du travail méthodologique que j'ai effectué, j'ai proposé 3 méthodes. Les profils taxonomiques résultants de ces méthodes tirent parti de l'universalité et du pouvoir discriminant de chaque marqueur, résultant en un profil taxonomique plus proche de la réalité. Que ce soit en terme d'identification des taxa ou d'estimation de leurs abondances relatives, les méthodes multi-marqueurs permettent d'obtenir de meilleurs résultats que celles n'utilisant qu'un seul marqueur. Les 3 méthodes multi-marqueurs donnent des résultats sensiblement identiques entre elles. Enfin nous avons remarqué que toutes les approches retrouvent la structure en 5 (ou 6) clusters mais que la faible résolution des approches mono-marqueurs ne permet pas systématiquement d'identifier l'espèce de *Lactobacillus* dominante et donc de faire la correspondance entre cluster et CST, contrairement aux approches multi-marqueurs.

3.5.2 . Recommandations d'usages

Bien que notre approche multi-marqueurs ne soit pas encore mature et nécessite des validations biologiques pour pouvoir être diffusée, il m'est déjà possible de formuler quelques préconisations d'usage.

Choix des marqueurs Il va de soi que, pour obtenir un profil taxonomique consensus satisfaisant, il faut que les marqueurs choisis soient pertinents pour l'écosystème étudié,qu'ils aient une résolution taxonomique globalement suffisante, un biais du nombre de copies limité, et que les bases de données associées soient les plus complètes possibles. Il ressort également de notre analyse qu'il vaut mieux être parcimonieux dans le choix des marqueurs. En effet, multiplier les marqueurs, s'ils n'apportent pas d'information complémentaire par rapport aux autres, ne semble pas bénéfique, notamment en raison des faux positifs qu'ils pourraient introduire.

Pipeline d'analyse pour chaque marqueur En amont de la méthode pour obtenir le consensus, le traitement des données pour chaque marqueur joue un rôle très important. Il est nécessaire de s'assurer que les profils taxonomiques issus des différents marqueurs partagent la

même taxonomie, c'est-à-dire les mêmes rangs taxonomiques et mêmes noms de taxa. Dans mon cas, j'ai du transposer les bases de données que j'avais choisies pour chaque marqueur à la taxonomie à jour du NCBI. Cette tâche, plus ou moins complexe selon les bases de données et les informations qui étaient à disposition pour chaque séquence, doit être réalisée avec précaution. Ensuite, lors de l'assignation taxonomique, il vaut mieux être conservateur, c'est à dire retourner un rang taxonomique plus élevé, afin d'éviter les faux positifs et compter sur les marqueurs réellement discriminants pour améliorer la résolution.

Limiter le profil d'un marqueur à une clade Il est possible que certains marqueurs soient utilisés dans le seul but d'améliorer la résolution taxonomique pour une clade en particulier. Si un marqueur est particulièrement bon pour une clade, mais mauvais pour le reste, il peut être intéressant de limiter le profil taxonomique à cette seule clade [7]. En effet, si ce marqueur est amplifié seulement chez certaines des bactéries des autres clades, les abondances de celles-ci seront fortement biaisées par l'universalité des amorces et pourrait dégrader l'estimation des abondances relatives dans les autres clades lors du calcul du profil consensus.

Choix de la méthode pour calculer le consensus J'ai développé trois méthodes différentes pour calculer le profil consensus, sans réussir à en identifier une qui serait systématiquement meilleure que les autres dans les tests que j'ai réalisés. Si les validations biologiques à suivre pourront peut-être éclairer ce choix, il existe tout de même quelques considérations à prendre en compte. Si les différents marqueurs n'ont pas la même universalité pour l'écosystème étudié, avec par exemple un marqueur qui est spécifique d'une clade, la méthode 2 ne devrait pas donner les meilleurs résultats. Dans le cas décrit dans la figure 3.1 page 54, cette méthode n'est pas adaptée et devrait donner des résultats qui ne sont pas conformes à la réalité. Cependant, nous n'avons pas observé dans nos résultats que cette méthode se comportait moins bien. Au contraire, elle donnait de meilleurs résultats (table 3.7, page 77). Nous avions pourtant des marqueurs plus universels que d'autres. Les méthodes 1 et 3 donnent des résultats comparables et peuvent toutes deux utiliser un système de pondération, il est donc difficile de mettre en avant l'une plus que l'autre.

3.5.3 . Limites et perspectives

Si la combinaison de gènes marqueurs semble prometteuse au vu de nos premiers résultats, il est certain que la méthode proposée au moment de la rédaction du présent manuscrit gagnerait à être améliorée dans les directions présentées ci-dessous.

Validation de la méthode Si les simulations que j'ai réalisées étaient nécessaires pour développer et tester la méthode, elles ne sont pas complétement conformes à des métagénomes réels. D'abord, le simulateur utilisé ne tolère aucun *mismatch* entre la séquence à amplifier et les amorces, alors que ces *mismatchs* sont possibles dans la réalité. Cela a pour conséquence de sur-estimer le biais d'universalité dans les jeux de données simulés. Par ailleurs, certains génomes à partir desquels sont simulés les *reads* ne sont pas complets. Ainsi, lorsqu'un marqueur n'est pas amplifié *in silico* dans un génome, l'absence d'amplfication du gène peut correspondre à un défaut d'assemblage du génome plutôt qu'à un défaut d'universalité des amorces.

Les méthodes, tant le pipeline bioinformatique qui analyse les données de chaque marqueur que la combinaison des profils taxonomiques pour obtenir le consensus, ont évidement besoin

7. avec la fonction `subset_taxa` de `phyloseq` ou équivalent

d'être mis à l'épreuve de données réelles. Il était prévu de produire à Alphabio un jeu de données d'une dizaine de prélèvements vaginaux, ainsi que des communautés bactériennes synthétiques de composition connue pour faire une première validation, avant d'envisager de plus gros jeux de données, avec plusieurs conditions cliniques. Les données n'ont malheureusement pas été produites au moment de la rédaction de ce manuscrit mais sont espérées prochainement.

Améliorations de l'assignation taxonomique pour chaque marqueur L'étape d'assignation taxonomique est perfectible en plusieurs points. D'abord, la transposition de la base de données de référence à la taxonomie du NCBI, étape nécessaire pour s'assurer que les profils taxonomiques issus des différents marqueurs partagent la même taxonomie, est perfectible. Pour certaines bases de données, cette transposition se fait par la recherche des noms d'espèces dans la taxonomie, et cela conduit dans certains cas à des approximations, voire à l'absence de résultats. Les sources de problèmes sont divers : il y a des changement de noms qui ne sont pas reconnus par les synonymes du NCBI, des noms mal orthographiés, des espèces inexistantes, et des cas compliqués comme "*environmental samples*" ou "*uncultured bacteria*". Si j'ai automatisé cette tâche avec un script python, utilisant `taxonkit` et `Bio.Entrez`, il pourrait être avantageux de vérifier et d'améliorer manuellement cette transposition, sur une sous-partie des séquences qui posent problème.

Par ailleurs, des espèces bactériennes présentes dans le microbiote vaginal sont manquantes dans certaines bases de données. On peut citer notamment le cas de BVAB1, une bactérie non cultivée mais retrouvée dans certains cas de vaginoses bactériennes, qui est absente de toutes les bases de données. Malgré l'absence d'isolat cultivé, on dispose pour cette espèce d'un génome complet issu de métagénomique, duquel on pourrait extraire les gènes marqueurs pour compléter les bases de données.

Enfin, lors de l'assignation taxonomique, j'ai mis au point un script pour traiter de manière automatique les résultats de l'alignement des séquences sur les bases de données, et gérer automatiquement les multi-affiliations. Cette automatisation n'est pas sans risque et les profils taxonomiques résultants ne sont pas optimaux. Entre autres, il semblerait important de définir des seuils de pourcentage d'identité minimal pour l'assignation et ainsi éviter la mauvaise sur-classification, c'est à dire la classification au niveau espèce alors que les informations ne permettent de descendre de façon fiable qu'au niveau genre, lorsque des taxa sont absents des bases de données, comme mentionné en page 79. La curation manuelle des alignements proposée dans FROGS permettrait une classification optimale et une meilleure confiance en les résultats.

Identification des souches L'identification taxonomique à un niveau plus fin que l'espèce est permis par certains marqueurs. Lorsque la base de données est suffisamment fournie, la diversité au sein des régions amplifiées peut permettre l'identification des sous-espèces. Cet aspect, particulièrement intéressant dans de nombreuses situations biologiques, est un objectif raisonnable pour notre approche multi-marqueurs. Néanmoins, une curation manuelle, tant des transpositions de la taxonomie des bases de données que des assignations taxonomiques, semble nécessaire pour assurer la fiabilité des résultats. Une approche phylogénétique, de type MLST (pour *multi-loci strain typing*), pourrait éventulellement appliquée, au cas par cas, si les marqueurs utilisés le permette. Ces approches sont classiquement utilisées pour caractériser les souches isolées (séquençage Sanger), en utilisant plusieurs marqueurs phylogénétiques spécifiques à certaines espèces ou groupes d'espèces.

Biais du nombre de copies Le biais du nombre de copies est un élément important qui explique les différences entre les profils taxonomiques issus de chaque marqueur. Comme observé dans la table 3.7 page 77, le fait de combiner les marqueurs permet de corriger, au moins partiellement, le biais du nombre de copies, en utilisant une moyenne des marqueurs. Des méthodes existent pour corriger en amont le biais du nombre de copies mais semblent toutes souffrir de faiblesses comme montré dans [96, 167]. Cependant, dans les cas des profils 16S et ITS_{bact}, qui possèdent le même biais du nombre de copies, il serait certainement intéressant de d'utiliser l'une de ces méthodes de correction avant de faire le consensus pour vérifier si elles améliorent réellement le résultat.

Pondération Le système de pondération proposé jusqu'ici, qui pénalise les marqueurs ayant des *reads* assignés au parent mais à aucun enfant, ne prend en compte que le pouvoir discriminant des marqueurs. Il serait intéressant de trouver un système de pondération qui favorise les marqueurs les plus universels. Par exemple, lorsqu'on cherche à établir la composition consensus au niveau des phyla, les disparités entre les profils sont davantage dues au biais d'universalité (et au biais du nombre de copies) qu'au manque de résolution taxonomique. En effet, certains marqueurs dont les amorces sont peu universelles, peuvent n'être amplifiés que chez une partie des bactéries appartenant à un phylum, sous-estimant ainsi son abondance relative. Il serait alors pertinent de donner moins de poids à ce marqueur qu'à un autre qui serait plus universel. Pour cela, il faudrait mettre au point un indicateur numérique qui, à partir des profils taxonomiques issus de différents marqueurs, évalue l'universalité de ces derniers au sein d'une clade. Il n'est cependant pas simple de mettre au point un tel indicateur numérique.

Processus récursif pour obtenir le profil consensus Le mode de fonctionnement adopté pour faire la combinaison, qui consiste à construire récursivement le profil taxonomique en commençant par le rang taxonomique le plus élevé avant de descendre vers le rang le plus bas, a permis de remplir l'objectif de produire un profil taxonomique consensus qui tire parti de l'universalité et du pouvoir discriminant des différents marqueurs. En revanche, il donne beaucoup d'importance à l'estimation des abondances relatives des rangs taxonomiques élevés, notamment des phyla, car tous les rangs inférieurs en dépendent. Des erreurs commises à un rang élevé vont donc être propagées vers les rangs plus bas. Or, les méthodes et les pondérations envisagées pour estimer les abondances relatives des enfants sont plus efficaces à des rangs taxonomiques bas. En effet, il est plus simple de trouver un consensus si les espèces bactériennes au sein de chaque enfant sont homogènes (par exemple, soit elles sont toutes amplifiées, soit aucune ne l'est). Lorsqu'on se place au niveau des phyla, chaque enfant contient une grande diversité d'espèces bactériennes hétérogènes, dans lesquelles se mélangent les biais d'universalité, les disparités de niveau d'assignations taxonomiques, les biais du nombres de copies, *etc*. Il est alors difficile d'établir un consensus qui ne se limite pas à une simple moyenne des valeurs issues des différents marqueurs.

Optimisation Les consensus ont jusqu'ici été calculés à partir des 5 marqueurs à chaque fois. Il serait intéressant de regarder si un sous-ensemble de ces marqueurs permettrait de calculer un meilleur consensus. Au delà du gain économique que cela représente, il est possible que les résultats soient meilleurs avec moins de marqueurs. En effet, en raison des faux positifs qui sont difficiles à éliminer au moment du consensus, rajouter des marqueurs qui n'apportent pas de nouvelle information n'est pas nécessairement bénéfique. Concrètement, on pourrait

comparer les résultats obtenus par consensus de tous les sous-ensembles de marqueurs, car même si la combinatoire est importante, les temps de calculs sont relativement faibles. De plus, conformément aux scénarios d'utilisation mentionnés en introduction de ce chapitre, il serait intéressant de voir l'impact de la profondeur de séquençage des différents marqueurs sur les résultats. Il serait notamment intéressant de voir si avoir un séquençage principal avec une grande couverture et un ou plusieurs séquençages satellites à faible profondeur donnerait de bons résultats.

Prédiction du profil fonctionnel La prédiction de profil fonctionnel à partir des profils taxonomiques obtenus par métabarcoding est une analyse que l'on retrouve dans beaucoup d'études, bien que cette projection souffre de nombreuses limitations dues à la faible conservation phylogénétique de certaines fonctions. Les outils permettant de réaliser cette projection [12, 42] sont également limités par la faible résolution taxonomique et la mauvaise quantification des abondances de certains profils. En effet, plus les bactéries et leurs abondances sont identifiées précisément, meilleures sont les prédictions de la composition en gènes du métagénome. Ainsi, on peut supposer que l'amélioration de la résolution des profils taxonomiques apportée par la combinaison des marqueurs est susceptible d'améliorer les profils fonctionnels résultants de ces outils. Il serait intéressant de confirmer cette hypothèse.

Autres écosystèmes Si cette méthode n'a pour l'instant été testée que dans le contexte du microbiote vaginal, il serait intéressant d'étudier sa pertinence dans d'autres écosystèmes, dont la composition est moins particulière. Dans un premier temps, il pourrait être intéressant d'utiliser les jeux de données publics qui présentent plusieurs marqueurs, pour lesquels j'ai déjà préparé la base de données de référence [137] et ceux qui contiennent plusieurs régions du 16S. Par ailleurs, on pourrait évaluer les performances des méthodes multi-marqueurs sur les jeux de données publics sur les eucaryotes mentionnés en introduction de ce chapitre [168, 38, 35, 174].

Valorisation Enfin, après validations biologiques et potentiels ajustements, j'aimerais partager la méthode sous forme de package R, et publier un article la présentant, ainsi que son intérêt dans le cadre du microbiote vaginal.

4 - APPLICATIONS CLINIQUES

Ce chapitre contient les deux articles publiés au cours de ma thèse qui ont été menés dans le cadre de ma collaboration avec Alphabio. Dans ces deux articles, mon rôle a été de conduire l'analyse bioinformatique et statistique, la visualisation et l'interprétation des données. Ces projets ont été menés en utilisant le métabarcoding ciblant la région V3V4 du gène 16S. L'analyse bioinformatique a été réalisée avec un pipeline que j'avais développé lors de mon stage de master, sous l'encadrement du Dr Anne Plauzolles à Alphabio et du Dr Ghislain Bidaut au CRCM. Dans ces deux projets, j'avais la responsabilité de la conception et de la réalisation des analyses statistiques et des figures, pour répondre aux questions biologiques posées. Je n'ai en revanche pas été impliqué dans la conception du plan expérimental en amont des projets. L'expertise biologique sur ces projets étaient portée par mes collaborateurs, j'ai été impliqué principalement dans la rédaction des sections correspondants au matériel et méthodes, ainsi que des résultats. Je remercie ici mes collaborateurs à Alphabio et à l'Hôpital Européen de Marseille, pour la confiance qui m'a été accordée et pour l'ensemble des travaux de recrutement des patients, de manipulations en laboratoire et d'études bibliographiques, qui m'ont permis de travailler sur ces jeux de données dans les meilleures conditions.

4.1 . *Human Stool Preservation Impacts Taxonomic Profiles in 16S Metagenomics Studies*

Le manque de consistance entre les résultats obtenus par différentes études portant sur le même sujet est un problème important dans l'étude des liens entre le microbiote humain et la santé. Le défaut de standardisation des protocoles utilisés pour analyser les microbiotes peut expliquer en partie ce phénomène. Bien sûr, ce n'est pas la seule source de variabilité des résultats entre les études, qui peut également être expliquée par la variabilité intrinsèque de la composition des microbiotes d'une population à une autre par exemple, la faible taille des cohortes, les choix de stratégie de séquençage et d'analyse des données. Les choix méthodologiques de collection et de conditionnement des échantillons (prélèvements de selles) avant l'analyse sont primordiaux et avaient déjà identifiés comme source de biais importants dans des études antérieures à notre projet.
Les méthodes de référence pour préserver le plus fidèlement l'échantillon sont d'extraire l'ADN immédiatement après la défécation, ou bien de congeler l'échantillon jusqu'à l'extraction de l'ADN. Cependant ces méthodes sont difficiles à mettre en place d'un point de vue pratique si l'on se projette dans une utilisation à grande échelle de l'analyse du microbiote. En effet, pour des raisons évidentes, il est plus confortable pour le patient de pouvoir réaliser le prélèvement à domicile que de devoir se déplacer en laboratoire pour le réaliser à une horaire donnée. Il également difficile de demander au patient de réaliser la congélation avec son congélateur domestique, pour des raisons d'hygiène, puis d'organiser le transport en assurant la chaîne du froid. Des solutions stabilisantes ont été développées pour permettre la conservation de l'échantillon à température ambiante. L'intérêt de cette étude était de comparer 10 solutions, ce qui est l'étude la plus exhaustive à notre connaissance, à travers un protocole expérimental particulièrement rigoureux de 14 jours incluant des variations de température.
Nous avons évalué les compositions de chaque échantillon par métabarcoding, et comparé les échantillons qui avaient subi le protocole expérimental à ceux qui avaient été congelés immédiatement, alors défini comme références. Nous avons constaté de très grands écarts dans la performance de stabilisation entre les solutions, et identifié les solutions qui marchaient le mieux. En outre, nous avons observé que certaines solutions, pourtant très utilisées et depuis de

longues années, marchent moins bien que de l'eau pour préserver la composition des échantillons dans notre protocole expérimental. Par ailleurs, nous avons montré que les abondances relatives des bactéries partageant certains traits phénotypiques (la coloration Gram et la tolérance à l'oxygène) étaient impactées de manière similaire par les différentes solutions stabilisantes. Ce lien entre phénotypes et altération était un résultat nouveau et intéressant pour comprendre les enjeux de la stabilisation des échantillons.
Si aucune solution ne préservait parfaitement la composition des échantillons à travers ce protocole particulièrement exigeant, on a montré que certaines étaient largement meilleures que d'autres, ce qui nous a permis de définir des préconisations pour un protocole standardisé de conservation des échantillons de selles.

frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 08 February 2022
doi: 10.3389/fcimb.2022.722886

Human Stool Preservation Impacts Taxonomic Profiles in 16S Metagenomics Studies

Anne Plauzolles[1†], Eya Toumi[1,2†], Marion Bonnet[1], Guillaume Pénaranda[1], Ghislain Bidaut[3], Laurent Chiche[4], Jérôme Allardet-Servent[5], Frédérique Retornaz[4], Benoit Goutorbe[1,3,6‡] and Philippe Halfon[1,4‡]*

[1] Clinical Research and R&D Department, Laboratoire Européen Alphabio, Marseille, France, [2] MEPHI, IHU Méditerranée Infection, Aix Marseille Université, Marseille, France, [3] CRCM, Aix-Marseille Univ U105, Inserm U1068, CNRS UMR7258, Institut Paoli-Calmettes, Marseille, France, [4] Infectious and Internal Medicine Department, Hôpital Européen Marseille, Marseille, France, [5] Intensive Care Unit, Hôpital Européen Marseille, Marseille, France, [6] Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

Microbiotas play critical roles in human health, yet in most cases scientists lack standardized and reproducible methods from collection and preservation of samples, as well as the choice of omic analysis, up to the data processing. To date, stool sample preservation remains a source of technological bias in metagenomic sequencing, despite newly developed storage solutions. Here, we conducted a comparative study of 10 storage methods for human stool over a 14-day period of storage at fluctuating temperatures. We first compared the performance of each stabilizer with observed bacterial composition variation within the same specimen. Then, we identified the nature of the observed variations to determine which bacterial populations were more impacted by the stabilizer. We found that DNA stabilizers display various stabilizing efficacies and affect the recovered bacterial profiles thus highlighting that some solutions are more performant in preserving the true gut microbial community. Furthermore, our results showed that the bias associated with the stabilizers can be linked to the phenotypical traits of the bacterial populations present in the studied samples. Although newly developed storage solutions have improved our capacity to stabilize stool microbial content over time, they are nevertheless not devoid of biases hence requiring the implantation of standard operating procedures. Acknowledging the biases and limitations of the implemented method is key to better interpret and support true associated microbiome patterns that will then lead us towards personalized medicine, in which the microbiota profile could constitute a reliable tool for clinical practice.

Keywords: microbiota, standardization, 16S metagenomics, human gut, preservation, stool, stabilizing solution.

OPEN ACCESS

Edited by:
Zhenjiang (Zech) Xu, Nanchang University, China

Reviewed by:
Lixin Luo, South China University of Technology, China
Angelica Cibrian-Jaramillo, Instituto Politécnico Nacional de México, Mexico

***Correspondence:**
Anne Plauzolles
a.plauzolles@alphabio.fr

†These authors have contributed equally to this work. Author order was determined in order of decreasing seniority

‡These authors have contributed equally to this work. Author order was determined in order of increasing seniority

Specialty section:
This article was submitted to Microbiome in Health and Disease, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *09 June 2021*
Accepted: *13 January 2022*
Published: *08 February 2022*

Citation:
Plauzolles A, Toumi E, Bonnet M, Pénaranda G, Bidaut G, Chiche L, Allardet-Servent J, Retornaz F, Goutorbe B and Halfon P (2022) Human Stool Preservation Impacts Taxonomic Profiles in 16S Metagenomics Studies. Front. Cell. Infect. Microbiol. 12:722886. doi: 10.3389/fcimb.2022.722886

INTRODUCTION

Over the past decade, an increasing number of studies have been published focusing on the human microbiome. While there is no doubt that these findings have helped us better comprehend the complexity of our microbiome and its implications on our health, the scientific community must compose with studies sometimes showing contradictory findings most often due to variabilities in protocols, cohorts' characteristics and sizes. The lack of standards hampers our expertise, as studies show inconsistencies, often resulting from technological bias rather than a true biological signature. To utilize microbiome science to its full potential, technical and computational methods must be standardized, and quality controls must be implemented to transition in the near future from a basic research environment to the clinic.

It is now common knowledge that our microbiome colonizes all body surfaces, especially our gut microbiome, which strongly impacts nearly every aspect of host physiology (Sekirov et al., 2010; Lozupone et al., 2012; Sommer and Bäckhed, 2013). Multiple lines of evidence now link alterations in the gut microbiome to numerous diseases (Ley et al., 2006; Sartor, 2008; Wen et al., 2008; Sekirov et al., 2010; Cryan and Dinan, 2012; Louis et al., 2014). However, microbiome studies most often lead to mixed results, halting our progress and hindering potential diagnosis, disease prediction and therapeutic intervention of microbiome analyses. Hence, most individual bacteria are not consistently associated with a given disease. Such discrepancies, in regard to microbiome signature patterns, are likely due to heterogeneity across study populations (small size, genetic factors, lifestyle) or the studied model or could be influenced by methodological differences among studies (Nguyen et al., 2015; Costea et al., 2017; Hornung et al., 2019).

The reality of microbiome research is that a variety of biological and technical factors can impact the quality of samples and their microbial content (Kim et al., 2017). The gut microbiome is the most challenging human ecosystem to characterize due to its heterogeneous bacterial populations. Its composition varies widely from one individual to another and involves a majority of bacterial populations that are very sensitive to oxygen (Conrads and Abdelbary, 2019), as well as remnants of human and food DNA and inhibitors likely to hamper subsequent analytical steps (Nechvatal et al., 2008). Technical bias can then result in misleading findings and can affect the quality of the data. Throughout the series of steps that a fecal sample undergoes to identify and characterize its microbial content, sampling and stabilization are key in the pre-analytical protocol and can heavily impact data quality. Previous studies have demonstrated that storage conditions of stool samples have only a small impact on their microbial content (Roesch et al., 2009; Lauber et al., 2010); however, more recent findings show otherwise (Cardona et al., 2012; Choo et al., 2015; Gorzelak et al., 2015; Guo et al., 2016; Hickl et al., 2019). DNA and RNA deteriorate rapidly after collection when kept at room temperature (Cardona et al., 2012), while the chemistry of existing stabilizing solutions has also demonstrated an impact on the recovery of genomic microbial content, resulting in a source of bias (Wu et al., 2019; Chen et al., 2020). Despite these conflicting results and challenges, a few principles are currently well acknowledged by the scientific community: avoid freeze-thaw cycles and temperature fluctuations throughout the preservation process (Cardona et al., 2012; Gorzelak et al., 2015; Thomas et al., 2015; Kim et al., 2017); when possible, shorten the transportation time; and freezing samples at -20°C or -80°C provides an optimal solution when immediate analysis of fresh sample is not an option (Wu et al., 2010; Bahl et al., 2012; Carroll et al., 2012; Fouhy et al., 2015; Gorzelak et al., 2015; Hale et al., 2015; Voigt et al., 2015; Shaw et al., 2016; Sinha et al., 2016; Song et al., 2016; Hickl et al., 2019; Wu et al., 2019).

In regard to studying microbiome composition using metagenomics, the method of collection that yields the most accurate results involves analyzing samples immediately after collection. However, this can be logistically challenging for samples such as stools that cannot be produced on demand. Any stabilizing method induces rapid changes in the presence and/or abundance of certain bacterial populations (Song et al., 2016). Despite different efficacies in stabilizing the true biological profile, the preservation step can result in biases even during short-term storage, but these alterations are, for most commonly utilized solutions, smaller or comparable to differences among technical replicates. Technical variability, albeit smaller than interindividual variability, may obscure subtle and meaningful alterations. Therefore, the choice of stabilization is highly dependent on factors such as limitations, availability, ease of use, cost and compatibility with the study's goals and/or 'omics' methods.

While the lack of standards affects the microbiome field in every 'omic' science and their related testing phases, including pre-analytical, analytical and post-analytical steps in sample processing, our research here focuses on technical bias in the pre-analytical handling of fecal samples in the study of gut microbiota through 16S metagenomics. For the past few years, the lack of standards and the sources of errors in datasets have been highlighted in the literature. Emerging protocols have arisen, but comparative studies, including comparison of most recent DNA stabilizers, are not sufficient to fully understand the bias that can emerge during this step. Our study aimed to evaluate and compare a large panel of stabilizing solutions that are either widely used by the scientific community or suited to the collection of fecal material. We also investigated the dynamic alterations that occurred over time in our samples based on their bacterial content and related phenotypical characteristics. Based on these results, acknowledging and identifying the limitations of DNA preservation could promote comparability among metagenomics studies and lead to clear guidelines that will be critical for scientific discovery going forward in understanding human microbiomes.

MATERIALS AND METHODS

Stool Collection and Ethic Approval

Fecal samples were collected from 15 French volunteers (11 women and 4 men) between 20 and 46 years old. Most samples

(n=10) were collected in the laboratory and handled immediately after defecation, while a minority (n=5) were collected at home and returned to our laboratory within 3 hours post-defecation. No medical records were collected. Age, gender and the sampling date were the only information provided by each volunteer.

All subjects provided written informed consent prior to participating in the study. These samples were anonymized and treated according to medical ethical guidelines.

Stool Conservation Study Design

To provide a standardized protocol for fecal sampling and preservation, fecal samples were collected from 15 French volunteers (n=15).

A total of 12 aliquots were evaluated over a period of 14 days. Ten aliquots were mixed with a stabilizer while two (one solid, one homogenized) were mixed with ultrasterile water and served as controls representatives of an unstabilized sample. Beforehand, fecal homogenization of the collected stool was performed. to limit variability among aliquots, allowing for better evaluation of the impact of the stabilizing solutions tested. In parallel, to evaluate the interaliquot variability, two sets of triplicates were compared, one set of solid stool aliquots and another set of homogenized stool aliquots. The immediate freezing of the homogenized set (Dc) allowed to preserve the microbial profile of the fecal sample at time of collection as freezing prevents the proliferation and deterioration of the microbial entities that compose a fecal sample. Hence an average of the Dc's triplicate was used as a reference profile in comparison to the aliquots preserved with a DNA stabilizer in order to evaluate the efficacy of the different stabilizers tested. Fecal homogenization was performed as follows: 12 g of stool was gently mixed with 30 ml of ultrasterile water for a few minutes. The collected stool was subsampled into 0.5 ml homogenized aliquots or 180-220 mg solid aliquots. All aliquots (solid and homogenized) were performed simultaneously and mixed with either a 1 ml of DNA stabilizer (for 10 aliquots, 1 to 10), or ultrasterile water (for two unstabilized controls, S and D) or immediately frozen at -20°C (for the two sets of triplicates, Sc and Dc). A total of 10 stabilizing solutions were tested: RNAlater (Ambion, Austin, US), Tris-EDTA (10 mM Tris-HCl pH 8.0, 1 mM EDTA) (Thermo Fisher Scientific, Massachusetts, US), 95% ethanol (VWR international, Pennsylvania, US), PrimeStore MTM (Longhorn Vaccines and Diagnostics, San Antonio, US), Stratec (Stratec Molecular GmbH, Berlin, Germany), OMNIgene-Gut (DNA Genotek, Ontario, Canada), Norgen (Norgen Biotek Corp., Thorold, Canada), DNA/RNA Shield (Zymo Research, Freiburg, Germany), Fecal Swab (Copan Italia S.P.A., Brescia, Italy), and Whatman FTA card (GE Healthcare Life Sciences, Illinois, US). For the FTA card method, a 0.5 ml sample was dispatched directly on the card. All stabilizers were tested on 15 fecal samples, except for PrimeStore MTM solution, which was tested on only 13 samples. All aliquots were preserved over a period of 14 days (**Figure 1**). Briefly, non-frozen aliquots (S, D and aliquots 1 to 10) were incubated for 14 days at varying temperatures fluctuating from 4°C to 40°C according to the following cycle: 3 days at room temperature (RT, approximately 25°C), 3 days at 4°C, 3 days at RT, 3 days at 40°C and 2 days at RT. These temperature fluctuations allowed evaluation of the efficacies of each stabilizing solution in harsh conditions. The temperature range chosen includes temperatures that a sample can be subjected to during transportation to the laboratory throughout the seasons for most countries worldwide.

Although, the collection of the fifteen samples was not uniform (either collected at home or in the laboratory), the reference for all sample was the profile at the time of stabilization. In this study, the alteration of all samples was compared to the reference samples (i.e., homogenized and immediately frozen samples stored with no additives, Dc).

Stool DNA Extraction

Bacterial DNA was isolated from all stool aliquots using the NucleoSpin® DNA Stool kit (Macherey-Nagel, Duren, Germany) following the manufacturer's instructions. Extracted DNA was stored at -20°C until subsequent application.

DNA Quantification and Purity Measurements

DNA quantification and purity (A260/A280 ratio) measurements were performed by spectrophotometry using a Nanodrop ND-1000 (Thermo Fisher Scientific, Massachusetts, US).

16S rRNA Gene Amplification, Library Preparation and High-Throughput Sequencing

To determine the bacterial composition of each aliquot, a 16S metagenomic sequencing library was created following Illumina's recommendations (Illumina, 2013). Briefly, this protocol targets the V3-V4 regions of the 16S rRNA gene during a first PCR using specific primers with overhang adapters: 16S Amplicon PCR Forward Primer 5' TCGTCGGCAGCGTCAGATGT GTATAAGAGACAGCCTACGGGNGGCWGCAG and 16S Amplicon PCR Reverse Primer 5' GTCTCGTGGGCT CGGAGATGTGTATAAGAGACAGGACTACH VGGGTATCTAATCC. Resulting amplicons were then purified using Agencourt AMPure XP magnetic beads (Beckman coulter, Brea, US). Subsequently, a second PCR was performed from the purified PCR amplicons to attach dual indices and Illumina sequencing adapters using the Nextera XT Index Kit (Illumina, San Diego, US). Following a second purification with Agencourt AMPure XP magnetic beads, the PCR products were then checked with quality controls using a fragment analyzer (Agilent Technologies, California, US) and Qubit (Thermo Fisher Scientific, Massachusetts, US) to evaluate DNA fragment sizes and DNA concentrations of the purified products. Barcoded amplicons were pooled in equal concentrations to generate a 4 nM library. The pool of samples was denatured to a final concentration of 12 pM and combined with 5% PhiX control (Illumina, San Diego, US). The 16S rRNA gene libraries were sequenced using a MiSeq instrument (Illumina, San Diego, US).

Experimental Validation

Demultiplexed and high-quality sequences (average quality score >Q30) were retrieved. Five samples with fewer than 30,000 reads were discarded, due to low quality DNA. All five samples were stabilized with Stratec Solution. A clustering

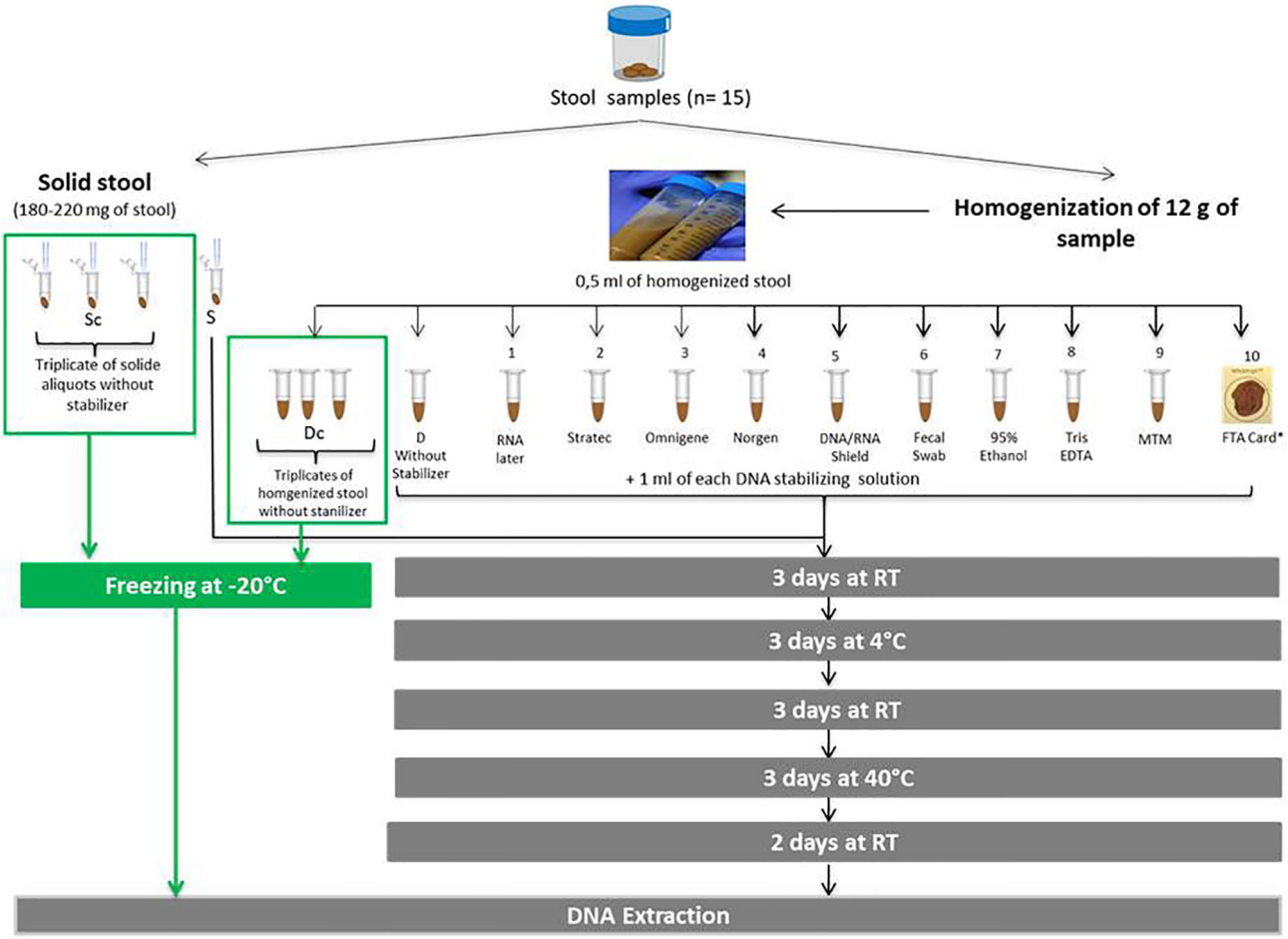


FIGURE 1 | Illustration of the experimental protocol. Evaluation of ten commercial DNA stabilizing solutions for the storage of fecal samples. (Sc, frozen solid sample; Dc, frozen homogenized sample; S, solid sample without stabilizer; D:, homogenized sample without stabilizer; RT, room temperature). * The whatman FTA card is a cotton-based cellulose matrix containing chemicals that lyse cells, denature proteins and protect DNA; 0.5 ml of homogenized solution was dispatched directly onto the card.

analysis was performed to validate the experiment. Three samples were excluded from our analysis as they did not cluster with their technological replicates (**Supplementary Figure S6** for details).

Bioinformatics Processing

Reads were processed using QIIME 2 (Bolyen et al., 2019) (version 2019.1.0) and its DADA2 (Callahan et al., 2016) plugin (q2dada2, version 2019.1.0). Preprocessing parameters were tuned to our dataset's specifications: reads were trimmed at their 3' ends at 245 bp, and reads shorter than this threshold were discarded; to remove amplification primers, 5' trimming was performed at 17 bp and 21 bp for forward and reverse reads, respectively. Reads that exceeded the 2 sequencing errors expected were discarded, and chimera removal was performed with the consensus method of DADA2. A denoising step was performed, and amplicon sequence variants (ASVs) were collected in a counting table. Taxonomic assignment was performed using Kraken (Wood and Salzberg, 2014) (version 1.1) based on the NCBI RefSeq Targeted Loci database, which contains over 21,000 bacterial and archaeal 16S reference sequences covering more than 15,000 species. Kraken approach, initially designed for shotgun metagenomics, was proved to be efficient for 16S reads when adapting the reference database used (Lu and Salzberg, 2020). Even if direct support for 16S databases was made available with Kraken 2, both versions share the same core concepts, with an enhancement of memory and time efficiency in the latest (Lu and Salzberg, 2020).

Statistical Analysis

Statistical analysis was performed with R (version 3.4) using the phyloseq package (McCurdie and Holmes, 2013) (version 1.22.3). Only ASVs with proportion beyond 10^{-4} in at least 5%

of the samples were considered for the analysis. Beta diversity was assessed with several metrics: Jaccard and Bray-Curtis dissimilarity indices were computed based on rarefied data (35 395 reads per sample were used for rarefaction), while Aitchison's distances (Gloor et al., 2017) were computed based on centered log-ratio transformed data with pseudo counts set at 0.5. Only the Bray-Curtis based analysis is shown, but different metrics confirmed this result (**Supplementary Figures S2–S4**).

First, the impact of homogenization was assessed by comparing mean distances within technological Sc (not homogenized) and Dc (homogenized) replicates across all stool samples collected using a paired Wilcoxon test.

Second, we evaluated stabilization performance by measuring the distance between each sample and its reference, defined as the barycenter of the 'Dc' replicates for the corresponding stool sample. We used a Kruskal-Wallis test to highlight the effect of the stabilizing solution on preservation of the bacterial content over storage time. Afterward, we performed a pairwise paired Wilcoxon test with Benjamini-Hochberg p-value correction for multiple hypothesis testing to determine which solutions performed better than others.

Finally, we searched for differentially abundant taxa, at the phylum and genus levels, between reference and stabilized samples with a Wilcoxon test, and p-values were adjusted with the Benjamini-Hochberg procedure. Furthermore, we gathered phenotypic data for the top 50 genera, representing up to 94% of all organisms found, regarding their oxygen sensitivity and their Gram stain status, as these characteristics are often conserved at the genus level (Schmaljohn and McClain, 1996; Brenner et al., 2005; Lowy, 2009). We used the LPSN database (Parte, 2018) to identify reference articles describing the characteristics of each genus. Gram stain status was defined as positive, negative or variable, and oxygen sensitivity was defined as strictly aerobic, strictly anaerobic, facultative anaerobic or microaerophile. Data are provided in **Supplementary Table T1**. We then clustered genera based on their median log2-fold change between stabilized samples and references using L2 distances and Ward's linkage to identify genera that behaved similarly in the stabilizing solutions tested. To track potential links between genus phenotype and behavior in stabilizing solutions, we performed a χ^2 test for independence of categorical variables between genera clusters and both oxygen sensitivity and Gram stain status independently. To further investigate these links, we aimed to determine whether phenotypic characteristics could prelude the emergence of the storage bias that we observed. Therefore, for each solution, we performed a Kruskal-Wallis test among genera for log2-fold change and both oxygen sensitivity and Gram stain status independently.

RESULTS

In the present study, the performance of each stabilizer was defined as the microbial community alterations over time relative to the reference sample (i.e., immediately frozen sample stored with no additive). The technical reproducibility of our analytical protocol was evaluated using triplicates of reference samples, while samples with no additive (S and D) served as indicators of the natural evolution of the microbiota profile if unstabilized.

Quality Control for DNA Yield, Purity and Alpha Diversity

Analysis of complex microbial ecosystems requires high-quality libraries for next generation sequencing (NGS) metagenomics. Hence, preserving a microbial profile over time and providing good DNA yield and purity of DNA extracts are key aspects in the analytical protocol in place. We found considerable differences in the DNA concentrations and A260/280 ratios of our extracted DNA. For example, Fecal Swab-preserved samples recovered, on average, 12-fold more DNA than Stratec-preserved samples (60.28 ng/µl vs 4.97 ng/µl). Among the different stabilizers tested in this study, recovered DNA was the lowest for Stratec, DNA/RNA Shield- and FTA card-stabilized samples (**Figure 2A**). In addition, samples preserved with these three solutions had primarily low A260/280 ratios (mean ratio <1.7), indicating the presence of contaminants (**Figure 2B**). Interestingly, Stratec samples were the least successful for recovering a microbiota profile with sufficient reads and showed a smaller alpha diversity than other stabilizers, while DNA/RNA Shield- and FTA card-preserved samples exhibited good profile recovery with high alpha diversity values (**Supplementary Figure S1**). The diversity index showed similar alpha diversity among preservation methods, and the Stratec stabilizer presented a much lower alpha diversity measure than the other stabilizers. As such, these results do not show any relationships among DNA concentration/purity, diversity and microbial profile recovery.

Dc and Sc measures resulted in similar concentrations and quality ratios among triplicate samples. Unstabilized samples (Dc, Sc, D and S) showed the highest DNA yield as they recovered on average 2.6-fold more DNA than stabilized samples with good A260/280 ratios.

A total of six samples were discarded, including five Stratec-stabilized samples and one Dc sample due to a lack of compliance with quality and/or quantity criteria.

Homogenization of Stool Samples Results in Reduced Intrasample Variability

Homogenization is commonly performed in studies to minimize intrasample variations and subsequent misestimation of the observed alterations within recovered profiles. The interaliquot variability for each Sc and Dc triplicate was first estimated by the mean distance using several methods (Bray-Curtis distance, Jaccard distance, and Aitchison distance). Comparison of distances within triplicates and between Sc and Dc triplicates showed a greater dispersion in Sc than in Dc triplicates, regardless of the distance method used (**Figure 3** and **Supplementary Figure S2**). In addition, a Wilcoxon test showed that homogenization significantly reduced observed interaliquot variability ($p=0.002$).

These results suggest that stool subsampling results in variations in the recovered microbial content among aliquots

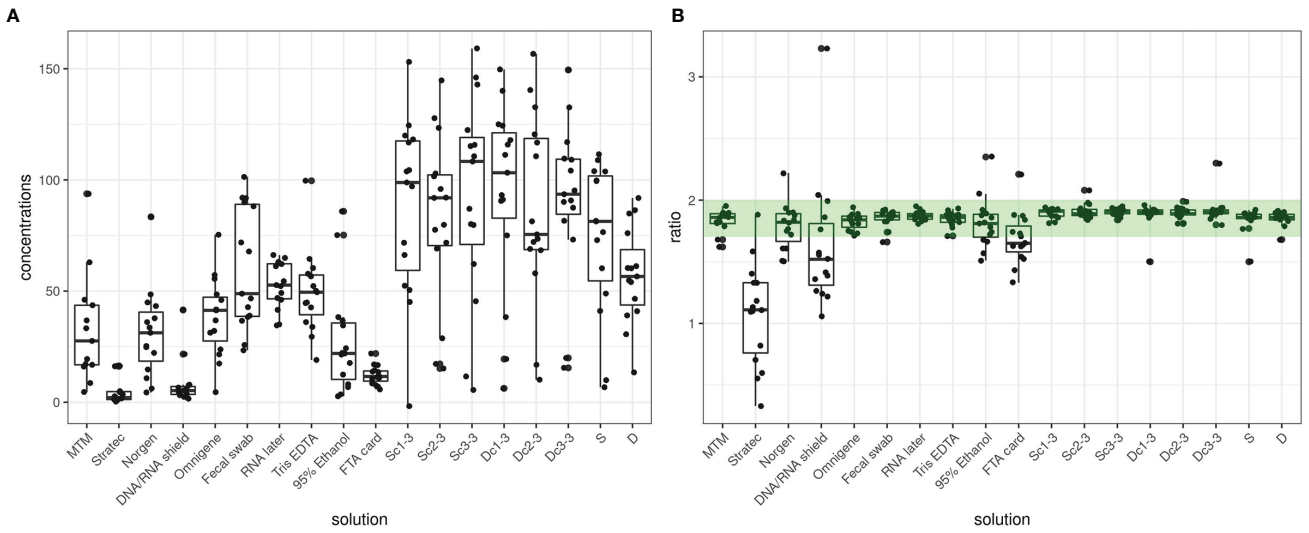


FIGURE 2 | Quantity and quality of DNA extracted from human fecal samples. **(A)** DNA yield, expressed as ng/µl. **(B)** A 260/280 nm ratio indicative of the presence or absence of phenol, solvent and protein-type contaminants in the DNA extract. The green range indicates a ratio between 1.7 and 2.0, which here defines an optimal DNA quality.

and confirms that homogenization of each sample has contributed here to significantly lowering the interaliquot variability. In this study, our homogenized aliquots added to the different stabilizers can thus be considered identical prior to storage. Their evolution over the 14-day storage period then provides an adequate evaluation of the efficacy of each stabilizer tested when compared to their reference (i.e., an average of Dc triplicates).

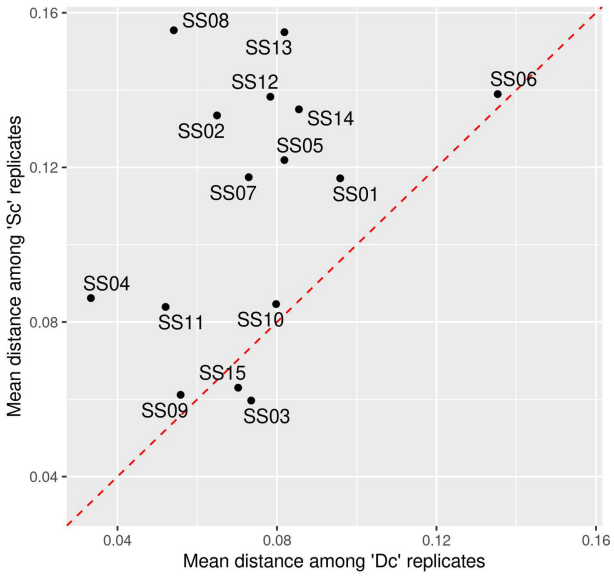


FIGURE 3 | Mean dispersion using Bray-Curtis distances among technological replicates prior (Sc) and after (Dc) homogenization across 15 fecal samples (SS01 to SS15). The red dotted line indicates equality of dispersion among 'Sc' and 'Dc' samples.

DNA Stabilizers Alter Stool Microbial Composition With Various Orders of Magnitude Compared to Samples With No Additives

To evaluate the performance of the tested stabilizers, we quantified the compositional dissimilarity between each preserved sample and its reference, defined as the barycenter of the 'Dc' replicates. Different metrics, including the Bray-Curtis, Jaccard, and Aitchison distances, show that Norgen, DNA/RNA Shield, OMNIgene-Gut and PrimeStore MTM produced profiles closest to their reference, while the remaining stabilizers resulted in greater alterations (**Figure 4A** and **Supplementary Figures S3A**, **S4A**). A Kruskal-Wallis test ($p<10^{-11}$) then confirmed that the solutions tested demonstrated distinct efficacies of stabilization specific to each stabilizer. Finally, a paired Wilcoxon test was used to compare the stabilizing performance among all stabilizers tested and identified Norgen as the best performing solution, closely followed by OMNIgene-Gut, DNA/RNA Shield and PrimeStore MTM, which presented similar performances (**Figure 4B** and **Supplementary Figures S3B**, **S4B**). In contrast, the least efficient stabilizers were Stratec, FTA card and Tris-EDTA, which appear no better than unstabilized samples (S or D).

In parallel, the results suggested that interindividual variability largely exceeded interaliquot variability (**Figure 4A** and **Supplementary Figures S3A**, **S4A**). Distances towards the reference were larger than interaliquot distances but smaller than those for interindividual variability indicating that preservation-induced effects were observed but were smaller than biological interindividual variability (**Supplementary Figure S5**). The only

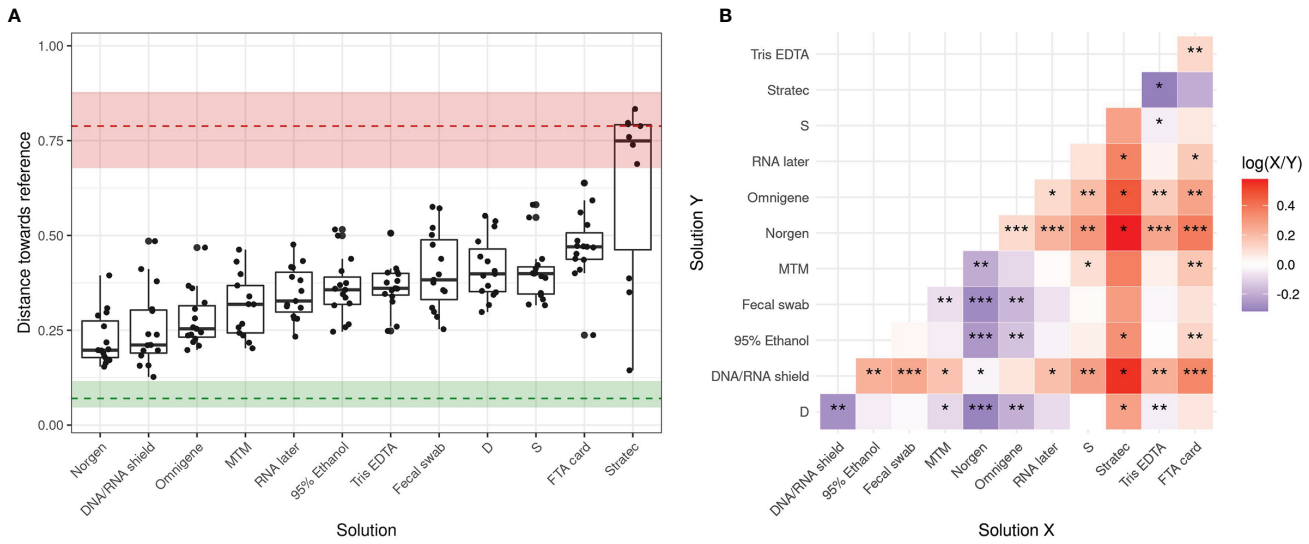


FIGURE 4 | Summary of community shifts in response to stabilizing solutions over a 14-day storage period. **(A)** Bray-Curtis distance towards the reference for each patient grouped by stabilizing solution. Median and 5th-to-95th percentile ranges are shown for both interaliquot and interpatient variability. **(B)** A pairwise paired Wilcoxon test was performed to compare solutions with each other, the color code refers to the log10 fold change of the median performance across patients: blue if the solution on X axis performs better, red if the solution on Y axis performs better. Significance is shown as follows: (*) indicates fdr<0.05, (**) indicates fdr<0.01, (***) indicates fdr<0.001.

exception was Stratec-preserved samples, which displayed a variability similar to that observed among samples, confirming that this solution is not suitable for storage of human fecal samples. These results were confirmed by hierarchical clustering as shown in **Supplementary Figure S6**.

Bacterial Relative Abundance Differs Based on the Method of Preservation in Different Taxonomic Ranks

Our analysis demonstrated that bacterial taxa were affected by the stabilizer, with misestimation of their relative abundance compared to their reference profiles (Dc). These alterations were detected at different taxonomic levels, including phyla (**Figure 5**) and genera (**Supplementary Figure S7**). Observed biases specific to each stabilizing solution were statistically confirmed by a paired Wilcoxon test, which showed that regardless of their efficiency of preserving a true microbiota profile, the different solutions tested impacted the relative abundance of certain bacterial taxa recovered when a fecal sample had been stored in a stabilizer. Low-abundance phyla (<1%), such as *Tenericutes*, *Synergistetes* and *Verrucomicrobia*, were the least significantly altered, except for *Lentisphaerae*, which was significantly overestimated in most storage conditions tested. Among abundant phyla (>1%), the most significantly affected were *Actinobacteria* and *Proteobacteria*, which tended to be overestimated, while *Firmicutes* and *Bacteroidetes* were underestimated. Of all abundant phyla, *Bacteroidetes* were interestingly the least significantly altered.

Parallel samples that were not exposed to any additive (S, D) also showed profile alterations, suggesting an effect of storage temporality, likely due to both bacterial growth for some populations and bacterial death for others. The lack of stabilization at fluctuating temperatures resulted in significant alterations of *Firmicutes*, *Proteobacteria*, *Actinobacteria* and *Lentisphaerae*.

Among the solutions with the greatest performances for stabilizing the fecal microbiota, Norgen did not significantly alter any phyla, except for *Lentisphaerae*, which were overestimated. In contrast, significant alterations were observed with OMNIgene-Gut and DNA/RNA Shield, both of which significantly affected *Firmicutes*, *Bacteroidetes*, *Proteobacteria* and *Lentisphaerae*, while *Actinobacteria* was only affected by DNA/RNA Shield. Interestingly, PrimeStore MTM appeared to significantly disturb only *Firmicutes* and *Lentisphaerae*. Stratec, which was the least efficient for preserving the fecal microbiota in our study, seemed to only affect *Actinobacteria*, but this result is biased, as many Stratec-preserved samples were excluded from this analysis due to poor quality DNA and/or low read numbers compared to the other stabilizing methods.

Considering the diversity of populations that can be found within phyla and the possibility that some phenotypic characteristics may dictate or facilitate certain alterations, it is interesting to observe changes at a lower taxonomic range. Genera clusters, based on alterations in the microbiota profile across all solutions tested, were found to be dependent on both oxygen sensitivity (p=0.0002) and Gram stain status (p=0.049) among the genera retrieved (**Supplementary Figure S7** and **Supplementary Table ST1**). These results indicate that the physiological traits examined could potentially prelude which populations are susceptible to alteration by the DNA stabilizers. For each

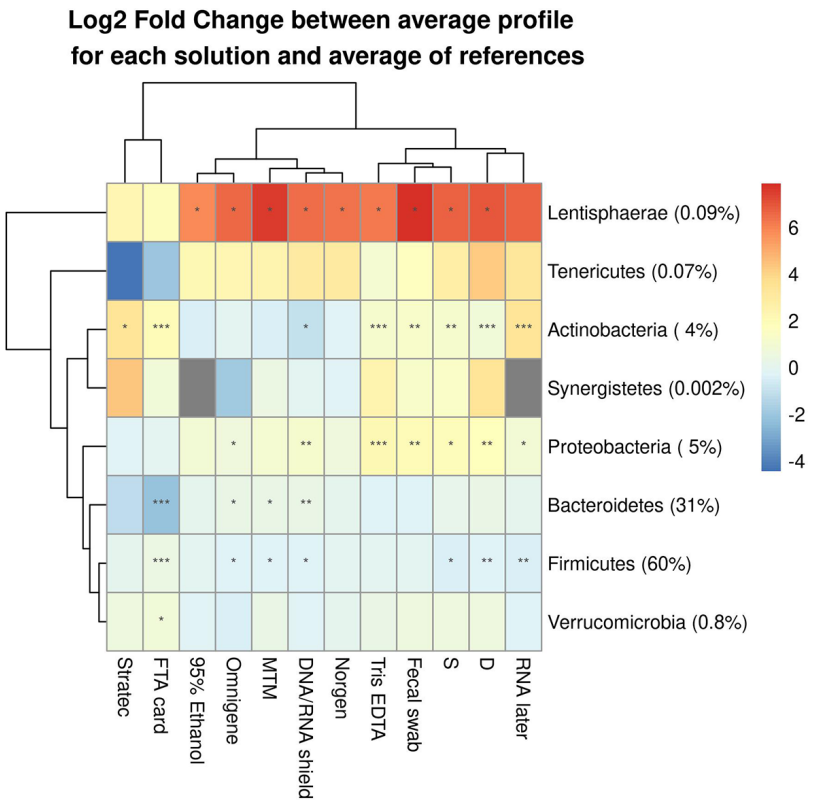


FIGURE 5 | Differentially abundant bacterial phyla between samples and their references among the 10 tested DNA stabilizing solutions. The median log2-fold change of average profiles is shown with the significance according the corresponding paired Wilcoxon test. (*) indicates fdr<0.05, (**) indicates fdr<0.01, (***) indicates fdr<0.001.

solution, we tested the effect of these phenotypical characteristics on the log2 fold change between a stabilized sample and its reference. We found no significant effect of oxygen sensitivity or Gram stain status on genus alterations in the absence of stabilizing solutions (samples S and D). In contrast, for stabilized samples, we found that genus alteration during storage was influenced by their oxygen status for Tris-EDTA (p=0.025) and in FTA card (p=0.029). Similarly, we found that Gram stain status affected samples stabilized with DNA/RNA Shield (p=0.002), PrimeStore MTM (p=0.027) and Stratec (p=0.043).

DISCUSSION

To the best of our knowledge, this study is the first in the microbiome field to compare such a large panel of storage methods, allowing identification of the best performing DNA stabilizers for a given ecosystem. We have shown that, of all stabilizers tested, some drastically impact the observed microbial composition and introduce biases. To proceed, we chose to evaluate methodologies already in use in the microbiome field through a comparative study of 10 storage methods to identify optimal fecal sampling methods that provide reproducible, stable, and accurate results.

Our analysis identified Norgen, OMNIgene-Gut, DNA/RNA Shield and PrimeStore MTM as the most efficacious stabilizers as compared to the immediately frozen aliquots (Dc). According to our results, several comparative studies have identified OMNIgene-Gut as a good DNA stabilizer for microbiome studies (Choo et al., 2015; Song et al., 2016; Abrahamson et al., 2017; Williams et al., 2019), while the other three solutions have not yet been extensively evaluated by comparative studies. In contrast, the remaining solutions tested were less efficient, showing a profile with alterations similar to unstabilized samples (S and D). Interestingly, among the stabilizers that were less reliable in our analysis, most showed discordant results in their ability to preserve fecal samples throughout comparative studies. For example, RNAlater was until recently the most commonly used buffer for metagenomic studies (Nechvatal et al., 2008; Cardona et al., 2012; Dominianni et al., 2014; Choo et al., 2015; Flores et al., 2015; Thomas et al., 2015; Song et al., 2016). However, its suitability for microbiome analysis has been extensively reviewed, as some studies claim that it results in reduced overall DNA yields and reduces the detection/abundance of bacterial taxa (Dominianni et al., 2014; Choo et al., 2015; Gorzelak et al., 2015; Hale et al., 2015; Sinha et al., 2016; Hickl et al., 2019). Our results did not show reduced DNA yield compared to other preserved solutions but did show

significant alterations in the recovered microbiota compared to their references, thus agreeing with previous studies that RNAlater is not an optimal preservation method. We came to the same conclusion for FTA card as Hale et al. (Hale et al., 2015), who demonstrated that FTA card (and RNAlater)-preserved samples were the least similar to fresh samples, while in contrast, Sinha et al. (Sinha et al., 2016) recommended the use of FTA card for short-term storage, demonstrating that it provides reproducible, stable, and accurate data across laboratories (over 4-day storage). The longer storage time in our protocol might have contributed to our discordant results. Similar to numerous studies performing homogenization of fecal samples (Carroll et al., 2012; Choo et al., 2015; Sinha et al., 2016; Song et al., 2016; Shaw et al., 2016; Vogtmann et al., 2017a; Vogtmann et al., 2017b), homogenization of our samples contributed to a better evaluation of the true performance of each stabilizing solution for preserving the microbiota content over time, as each aliquot presented a similar profile when added to the stabilizer.

Despite various effective methods for preserving a true microbiota profile over storage time, the alterations observed between the reference samples and their 14-day-stabilized aliquots were smaller than the differences between samples (subjects), except for Stratec-preserved samples. Furthermore, triplicates for each stool sample collected did not cluster by preservation method. Therefore, the human gut appears to be highly subject-specific, as our results suggest that interindividual variation accounts for the major of differences observed in fecal samples and outweighs the effect (or bias) of collection and storage, as previously demonstrated in several studies (Choo et al., 2015; Voigt et al., 2015; Guo et al., 2016; Sinha et al., 2016; Song et al., 2016). As stated above, the only exception was Stratec-preserved samples, which displayed variability similar to that observed among samples, indicating that this solution is not suitable for storage of human fecal samples. This result contradicts a recent study (Chen et al., 2020), which concluded that the Stratec solution was a suitable storage buffer for fecal specimen preservation. However, Chen et al. performed this study on a small cohort (n=4) over a 7-day period of storage at room temperature. The fluctuating temperatures in our protocol and the longer period of storage might explain the discrepancies between these findings. Additionally, our results did not demonstrate any relationships among DNA concentration/purity, microbial diversity, and microbial composition, similar to previous studies (Salonen et al., 2010; Hale et al., 2015). However, it has been suggested that high DNA concentrations might favor the identification of rare populations (Dominianni et al., 2014; Choo et al., 2015; Gorzelak et al., 2015; Hale et al., 2015). Although the low DNA yield observed with Stratec-stabilized samples might not entirely explain the difficulties in recovering a good microbiota profile, this factor may have contributed to its poor performances in our protocol.

Finally, microbiome comparative studies investigating the effect of storage often examine variations in the relative abundances of phyla and genera specific to the stabilizing methods. However, they do not examine these alterations based on microbial population characteristics, with the literature showing that bacteria within a genus share the same general phenotypic characteristics, in particular oxygen sensitivity and Gram stain status (Schmaljohn and McClain, 1996; Brenner et al., 2005; Lowy, 2009). In this study, we demonstrated that altered dynamics resulting from sample preservation are dictated by the phenotypical characteristics of the bacterial populations present in the studied sample. Our samples showed that genera alteration during storage is influenced by oxygen status for the Tris-EDTA and FTA card methods, as well as the Gram stain status for the DNA/RNA Shield, PrimeStore MTM and Stratec methods. A recent study also demonstrated that Gram status can alter the microbial content when Norgen stabilizer is used (Watson et al., 2019). Hence, preservation of the microbiota profile is impacted by the stabilizer chosen and its efficacy for preserving the true microbial profile. However, it must be taken into consideration that the stabilizer's performance can also be affected by the microbial content of the studied sample and its most common phenotypical traits.

One limitation of our study is that we did not evaluate the stabilizing performances of each solution tested across different times or over long-term storage periods. Indeed, Sinha et al. (Sinha et al., 2016) found that incubation at room temperature over 4 days reduced the reproducibility for most sampling methods, including no additives, swab, 70% ethanol, and EDTA. As such, the performance measures in our study only reflect their efficacies over a period of 14 days throughout various temperature fluctuations but do not attest of the loss of technical reproducibility or the impact on the alteration of bacterial taxa if samples are incubated in their stabilizers for a longer period. Considering we demonstrate here that genera alteration is influenced by oxygen status, our results may therefore underestimate the impact of the stabilizing solution on oxygen-sensitive bacterial population for the samples collected at home. Finally, our results here show that the choice of the stabilizer could be impacted by the microbial composition and phenotypic traits of the studied samples, but our study only analyzed the human gut ecosystem of a French cohort on a small group of individuals. Hence efficacies of the different DNA stabilizers used by the scientific community could vary depending on the microbial composition of the studied population. A study of a larger cohort with varying genetics (from different ethnic groups) and environmental factors could then lead to different conclusions and rank the efficacy of the DNA stabilizers differently. Further studies will be required in order to provide the scientific community with a more comprehensive analysis of stabilizing methods throughout different cohorts and with different types of samples to establish guidelines that will help scientists in their experimental settings.

We anticipate that procedures for microbial preservation will likely further improve in the future, and we show with this study that preservation remains a key step that can introduce technical bias into the study of complex ecosystems such as the human gut. Here, we demonstrated that some stabilizers are not suitable for the preservation of a stool sample when the sample is intended to

describe the whole complexity of the human gut ecosystem through 16S metagenomics. Our data identified Norgen, OMNIgene-Gut, DNA/RNA Shield and PrimeStore MTM as the most effective stabilizers, as they resulted in reduced technical biases. Acknowledging the performances of stabilizing solutions and their suitability depending on the microbial content of the ecosystem studied will help establish standards in omics studies. If implemented within metagenomics protocols across laboratories, these solutions could promote experimental reproducibility among research groups and lead to meaningful knowledge about the gut microbiome and its impact on human health with the discovery of new health-associated microbiome patterns and biomarkers.

CONCLUSION

The diversity and complexity of the human gut microbiota increase the difficulty of elaborating a method to study such ecosystems without experimental biases. Storage conditions can introduce substantial changes to microbial community profiling in regard to 16S metagenomics. Acknowledging the biases and limitations of the implemented method is key to better interpret and support true health (disease)-associated microbiome patterns that will then lead us towards personalized medicine, in which the microbiota profile could constitute a reliable tool for clinical practice.

DATA AVAILABILITY STATEMENT

The data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB40569 (https://www.ebi.ac.uk/ena/browser/view/PRJEB40569). Scripts are available at https://gitcrcm.marseille.inserm.fr/goutorbe/stool-preservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AP designed the study, extracted and sequenced samples, analyzed data, and wrote and finalized the manuscript. ET extracted and sequenced samples and drafted the manuscript. BG analyzed data and drafted the manuscript. MB extracted and sequenced samples. GP analyzed data. PH finalized the manuscript and funded this study. All authors discussed the results and commented on the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Alphabio laboratory funded this study.

ACKNOWLEDGMENTS

We thank our collaborators, Alphabio's molecular biology team, the CRCM Integrative Bioinformatics platform and Bernard Chetrit from the DataCentre for IT and Scientific Computing (Disc) platform for their support and involvement in the management of this study. We also thank Mahendra Mariadassou for his comments on the statistical analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2022.722886/full#supplementary-material

Supplementary Figure 1 | Effects of storage conditions on alpha diversity with respect to the observed richness and Shannon index.

Supplementary Figure 2 | Mean dispersion using the Jaccard distance **(A)** or Aitchison distance **(B)** among technological replicates prior to (Sc) and after (Dc) homogenization across 15 fecal samples. The red dotted lines indicate the equality of dispersion among 'Sc' and 'Dc' samples.

Supplementary Figure 3 | Summary of community shifts in response to stabilizing solutions over a 14-day storage period. **(A)** Jaccard distance towards the reference for each participant, grouped by stabilizing solution, and the median and 5th-to-95th percentile range are shown for both interaliquot and interpatient variability. **(B)** A pairwise paired Wilcoxon test was performed to compare solutions with each other, the color code refers to the log10 fold change of the median performance across patients: blue means that the solution on X axis performs better, red means that the solution on Y axis performs better. Significance is shown as follows: (*) indicates fdr<0.05, (**) indicates fdr<0.01, (***) indicates fdr<0.001.

Supplementary Figure 4 | Summary of community shifts in response to stabilizing solutions over a 14-day storage period. **(A)** The Aitchison distance towards the reference for each participant, grouped by stabilizing solution, median and 5th-to-95th percentile range are shown for both interaliquot and interpatient variability. **(B)** A pairwise paired Wilcoxon test was performed to compare solutions with each other, the color code refers to the log10 fold change of the median performance across patients: blue means that the solution on X axis performs better, red means that the solution on Y axis performs better. Significance is shown as follows: (*) indicates fdr<0.05, (**) indicates fdr<0.01, (***) indicates fdr<0.001.

Supplementary Figure 5 | Principal coordinate analysis (PCoA) based on Bray-Curtis distances computed from the rarefied data set. The plots are split according to samples origins, colored and labelled according to storage method. Points labelled as REF refers to the reference used for evaluation of storage methods for each patient, defined as the mean of the Dc samples.

Supplementary Figure 6 | Hierarchical clustering based on the Bray-Curtis distance matrix of all samples in the data set. The first 4 digits of the sample IDs refer to the biological origin of the fecal sample, and the remaining digits refer to the storage conditions (i.e., Sc, Dc, D, S or stabilizing solutions). Technological replicates clustered together, except for *SS01Sc1-3*, *SS02Sc1-3* and *SS04Dc3-3* (shown in red), which were excluded from downstream analysis.

Supplementary Figure 7 | Differentially abundant bacterial genera among samples and their references among 10 tested DNA stabilizing solutions. The median log2-fold change between average profiles and significance of the corresponding paired Wilcoxon test are shown. (*) indicates fdr<0.05, (**) indicates fdr<0.01, (***) indicates fdr<0.001.

Supplementary Table 1 | Phenotypic characteristics of the 50 most abundant genera recovered following a 14-day storage period. NEG, negative Gram strain; POS, positive Gram strain; VAR, variable Gram stain; ANAER-ST, strict anaerobic; AER-ST, strict aerobic; FAC, facultative anaerobic; MICRO-AE, microaerophile.* Proportion (%), average relative abundance of each genera within the overall dataset.

REFERENCES

Abrahamson, M., Hooker, E., Ajami, N. J., Petrosino, J. F., and Orwoll, E. S. (2017). Successful Collection of Stool Samples for Microbiome Analyses From a Large Community-Based Population of Elderly Men. *Contemp. Clin. Trials Commun.* 7, 158–162. doi: 10.1016/j.conctc.2017.07.002

Bahl, M. I., Bergström, A., and Licht, T. R. (2012). Freezing Fecal Samples Prior to DNA Extraction Affects the Firmicutes to Bacteroidetes Ratio Determined by Downstream Quantitative PCR Analysis. *FEMS Microbiol. Lett.* 329, 193–197. doi: 10.1111/j.1574-6968.2012.02523.x

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Brenner, D. J., Staley, J. T., and Krieg, N. R. (2005). "Classification of Procaryotic Organisms and the Concept of Bacterial Speciation," in *Bergey's Manual® of Systematic Bacteriology: Volume Two: The Proteobacteria, Part A Introductory Essays*. Eds. D. J. Brenner, N. R. Krieg, J. T. Staley and G. M. Garrity (Boston, MA: Springer US), 27–32. doi: 10.1007/0-387-28021-9_4

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-Resolution Sample Inference From Illumina Amplicon Data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., et al. (2012). Storage Conditions of Intestinal Microbiota Matter in Metagenomic Analysis. *BMC Microbiol.* 12, 1–8. doi: 10.1186/1471-2180-12-158

Carroll, I. M., Ringel-Kulka, T., Siddle, J. P., Klaenhammer, T. R., and Ringel, Y. (2012). Characterization of the Fecal Microbiota Using High-Throughput Sequencing Reveals a Stable Microbial Community During Storage. *PloS One* 7, e46953. doi: 10.1371/journal.pone.0046953

Chen, C.-C., Wu, W.-K., Chang, C.-M., Panyod, S., Lu, T.-P., Liou, J.-M., et al. (2020). Comparison of DNA Stabilizers and Storage Conditions on Preserving Fecal Microbiota Profiles. *J. Formosan. Med. Assoc.* 119 (12), 1791–1798. doi: 10.1016/j.jfma.2020.01.013

Choo, J. M., Leong, L. E., and Rogers, G. B. (2015). Sample Storage Conditions Significantly Influence Faecal Microbiome Profiles. *Sci. Rep.* 5, 1–10. doi: 10.1038/srep16350

Conrads, G., and Abdelbary, M. M. (2019). Challenges of Next-Generation Sequencing Targeting Anaerobes. *Anaerobe* 58, 47–52. doi: 10.1016/j.anaerobe.2019.02.006

Costea, P. I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., et al. (2017). Towards Standards for Human Fecal Sample Processing in Metagenomic Studies. *Nat. Biotechnol.* 35, 1069–1076. doi: 10.1038/nbt.3960

Cryan, J. F., and Dinan, T. G. (2012). Mind-Altering Microorganisms: The Impact of the Gut Microbiota on Brain and Behaviour. *Nat. Rev. Neurosci.* 13, 701–712. doi: 10.1038/nrn3346

Dominianni, C., Wu, J., Hayes, R. B., and Ahn, J. (2014). Comparison of Methods for Fecal Microbiome Biospecimen Collection. *BMC Microbiol.* 14, 103. doi: 10.1186/1471-2180-14-103

Flores, R., Shi, J., Yu, G., Ma, B., Ravel, J., Goedert, J. J., et al. (2015). Collection Media and Delayed Freezing Effects on Microbial Composition of Human Stool. *Microbiome* 3, 33. doi: 10.1186/s40168-015-0092-7

Fouhy, F., Deane, J., Rea, M. C., O'Sullivan, Ó., Ross, R. P., O'Callaghan, G., et al. (2015). The Effects of Freezing on Faecal Microbiota as Determined Using MiSeq Sequencing and Culture-Based Investigations. *PloS One* 10, e0119355. doi: 10.1371/journal.pone.0119355

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* 8. doi: 10.3389/fmicb.2017.02224

Gorzelak, M. A., Gill, S. K., Tasnim, N., Ahmadi-Vand, Z., Jay, M., and Gibson, D. L. (2015). Methods for Improving Human Gut Microbiome Data by Reducing Variability Through Sample Processing and Storage of Stool. *PloS One* 10, e0134802. doi: 10.1371/journal.pone.0134802

Guo, Y., Li, S.-H., Kuang, Y.-S., He, J.-R., Lu, J.-H., Luo, B.-J., et al. (2016). Effect of Short-Term Room Temperature Storage on the Microbial Community in Infant Fecal Samples. *Sci. Rep.* 6, 26648. doi: 10.1038/srep26648

Hale, V. L., Tan, C. L., Knight, R., and Amato, K. R. (2015). Effect of Preservation Method on Spider Monkey (Ateles Geoffroyi) Fecal Microbiota Over 8weeks. *J. Microbiol. Methods* 113, 16–26. doi: 10.1016/j.mimet.2015.03.021

Hickl, O., Heintz-Buschart, A., Trautwein-Schult, A., Hercog, R., Bork, P., Wilmes, P., et al. (2019). Sample Preservation and Storage Significantly Impact Taxonomic and Functional Profiles in Metaproteomics Studies of the Human Gut Microbiome. *Microorganisms* 7, 367. doi: 10.3390/microorganisms7090367

Hornung, B. V., Zwittink, R. D., and Kuijper, E. J. (2019). Issues and Current Standards of Controls in Microbiome Research. *FEMS Microbiol. Ecol.* 95, fiz045. doi: 10.1093/femsec/fiz045

Illumina (2013) *16s Sample Preparation Guide*. Available at: https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf (Accessed June 7, 2021).

Kim, D., Hofstaedter, C. E., Zhao, C., Mattei, L., Tanes, C., Clarke, E., et al. (2017). Optimizing Methods and Dodging Pitfalls in Microbiome Research. *Microbiome* 5, 1–14. doi: 10.1186/s40168-017-0267-5

Lauber, C. L., Zhou, N., Gordon, J. I., Knight, R., and Fierer, N. (2010). Effect of Storage Conditions on the Assessment of Bacterial Community Structure in Soil and Human-Associated Samples. *FEMS Microbiol. Lett.* 307, 80–86. doi: 10.1111/j.1574-6968.2010.01965.x

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Human Gut Microbes Associated With Obesity. *Nature* 444, 1022–1023. doi: 10.1038/4441022a

Louis, P., Hold, G. L., and Flint, H. J. (2014). The Gut Microbiota, Bacterial Metabolites and Colorectal Cancer. *Nat. Rev. Microbiol.* 12, 661–672. doi: 10.1038/nrmicro3344

Lowy, F. (2009). *Bacterial Classification, Structure and Function* (New York, USA: Columbia University), 1–6.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, Stability and Resilience of the Human Gut Microbiota. *Nature* 489, 220–230. doi: 10.1038/nature11550

Lu, J., and Salzberg, S. L. (2020). Ultrafast and Accurate 16S rRNA Microbial Community Analysis Using Kraken 2. *Microbiome* 8, 124. doi: 10.1186/s40168-020-00900-2

McMurdie, P. J., and Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PloS One* 8, e61217. doi: 10.1371/journal.pone.0061217

Nechvatal, J. M., Ram, J. L., Basson, M. D., Namprachan, P., Niec, S. R., Badsha, K. Z., et al. (2008). Fecal Collection, Ambient Preservation, and DNA Extraction for PCR Amplification of Bacterial and Human Markers From Human Feces. *J. Microbiol. Methods* 72, 124–132. doi: 10.1016/j.mimet.2007.11.007

Nguyen, T. L. A., Vieira-Silva, S., Liston, A., and Raes, J. (2015). How Informative is the Mouse for Human Gut Microbiota Research? *Dis. Models Mech.* 8, 1–16. doi: 10.1242/dmm.017400

Parte, A. C. (2018). LPSN - List of Prokaryotic Names With Standing in Nomenclature (Bacterio.Net), 20 Years on. *Int. J. Syst. Evol. Microbiol.* 68, 1825–1829. doi: 10.1099/ijsem.0.002786

Roesch, L. F., Casella, G., Simell, O., Krischer, J., Wasserfall, C. H., Schatz, D., et al. (2009). Influence of Fecal Sample Storage on Bacterial Community Diversity. *Open Microbiol. J.* 3, 40. doi: 10.2174/1874285800903010040

Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R. A., et al. (2010). Comparative Analysis of Fecal DNA Extraction Methods With Phylogenetic Microarray: Effective Recovery of Bacterial and Archaeal DNA Using Mechanical Cell Lysis. *J. Microbiol. Methods* 81, 127–134. doi: 10.1016/j.mimet.2010.02.007

Sartor, R. B. (2008). Microbial Influences in Inflammatory Bowel Diseases. *Gastroenterology* 134, 577–594. doi: 10.1053/j.gastro.2007.11.059

Schmaljohn, A. L., and McClain, D. (1996). "Alphaviruses (Togaviridae) and Flaviviruses (Flaviviridae)," in *Medical Microbiology*, *4th* Gavelston (University of Texas Medical Branch at Galveston).

Sekirov, I., Russell, S. L., Antunes, L. C. M., and Finlay, B. B. (2010). Gut Microbiota in Health and Disease. *Physiol. Rev.* 90, 859–904. doi: 10.1152/physrev.00045.2009

Shaw, A. G., Sim, K., Powell, E., Cornwell, E., Cramer, T., McClure, Z. E., et al. (2016). Latitude in Sample Handling and Storage for Infant Faecal Microbiota Studies: The Elephant in the Room? *Microbiome* 4, 40. doi: 10.1186/s40168-016-0186-x

Sinha, R., Chen, J., Amir, A., Vogtmann, E., Shi, J., Inman, K. S., et al. (2016). Collecting Fecal Samples for Microbiome Analyses in Epidemiology Studies. *Cancer Epidemiol. Biomarkers Prev.* 25, 407–416. doi: 10.1158/1055-9965.EPI-15-0951

Sommer, F., and Bäckhed, F. (2013). The Gut Microbiota—Masters of Host Development and Physiology. *Nat. Rev. Microbiol.* 11, 227–238. doi: 10.1038/nrmicro2974

Song, S. J., Amir, A., Metcalf, J. L., Amato, K. R., Xu, Z. Z., Humphrey, G., et al. (2016). Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems* 1, e00021–16. doi: 10.1128/mSystems.00021-16

Thomas, V., Clark, J., and Doré, J. (2015), e00021–16. Fecal Microbiota Analysis: An Overview of Sample Collection Methods and Sequencing Strategies. *Future Microbiol.* 10, 1485–1504. doi: 10.2217/fmb.15.87

Vogtmann, E., Chen, J., Amir, A., Shi, J., Abnet, C. C., Nelson, H., et al. (2017a). Comparison of Collection Methods for Fecal Samples in Microbiome Studies. *Am. J. Epidemiol.* 185, 115–123. doi: 10.1093/aje/kww177

Vogtmann, E., Chen, J., Kibriya, M. G., Chen, Y., Islam, T. , Eunes, M., et al. (2017b). Comparison of Fecal Collection Methods for Microbiota Studies in Bangladesh. *Appl. Environ. Microbiol.* 83, e00361–17. doi: 10.1128/AEM.00361-17

Voigt, A. Y., Costea, P. I., Kultima, J. R., Li, S. S., Zeller, G., Sunagawa, S., et al. (2015). Temporal and Technical Variability of Human Gut Metagenomes. *Genome Biol.* 16, 73. doi: 10.1186/s13059-015-0639-8

Watson, E.-J., Giles, J., Scherer, B. L., and Blatchford, P. (2019). Human Faecal Collection Methods Demonstrate a Bias in Microbiome Composition by Cell Wall Structure. *Sci. Rep.* 9, 1–18. doi: 10.1038/s41598-019-53183-5

Wen, L., Ley, R. E., Volchkov, P. Y., Stranges, P. B., Avanesyan, L., Stonebraker, A. C., et al. (2008). Innate Immunity and Intestinal Microbiota in the Development of Type 1 Diabetes. *Nature* 455, 1109–1113. doi: 10.1038/nature07336

Williams, G. M., Leary, S. D., Ajami, N. J., Chipper Keating, S., Petrosin, J. F., Hamilton-Shield, J. P., et al. (2019). Gut Microbiome Analysis by Post: Evaluation of the Optimal Method to Collect Stool Samples From Infants Within a National Cohort Study. *PloS One* 14, e0216557. doi: 10.1371/journal.pone.0216557

Wood, D. E., and Salzberg, S. L. (2014). Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. *Genome Biol.* 15, R46. doi: 10.1186/gb-2014-15-3-r46

Wu, W.-K., Chen, C.-C., Panyod, S., Chen, R.-A., Wu, M.-S., Sheen, L.-Y., et al. (2019). Optimization of Fecal Sample Processing for Microbiome Study—The Journey From Bathroom to Bench. *J. Formosan. Med. Assoc.* 118, 545–555. doi: 10.1016/j.jfma.2018.02.005

Wu, G. D., Lewis, J. D., Hoffmann, C., Chen, Y.-Y., Knight, R., Bittinger, K., et al. (2010). Sampling and Pyrosequencing Methods for Characterizing Bacterial Communities in the Human Gut Using 16S Sequence Tags. *BMC Microbiol.* 10, 206. doi: 10.1186/1471-2180-10-206

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4.2 . *Gut microbiota in systemic lupus erythematosus patients and lupus mouse model : A cross species comparative analysis for biomarker discovery*

La littérature scientifique présente aujourd'hui des preuves solides de l'implication du microbiote intestinal dans les fonctions immunitaires de l'hôte, et du fait que le dérèglement de l'immunité peut s'accompagner d'une altération de la composition du microbiote intestinal. Le lupus érythémateux systémique est une maladie auto-immune chronique caractérisée par la production d'anticorps anti-nucléaires s'attaquant à l'ADN natif qui touche plusieurs organes, avec notamment des manifestations cutanées, articulaires, hépatiques et intestinales. Les patients alternent souvent entre phases de rémission et phases de poussée de symtômes. Ce projet s'inscrit dans un vaste programme de science participative mené par le Dr Laurent Chiche à l'Hôpital Européen de Marseille, appelé 3L1 pour *Lupus Living Lab*, qui vise à suivre des patients lupiques sur plusieurs années pour évaluer les facteurs influençant les poussées de symptômes, pas seulement le microbiote, et améliorer la qualité de vie des patients.
En comparant la composition du microbiote intestinal de 16 patients lupiques très bien caractérisés cliniquement avec celle de 76 témoins asymptomatiques, nous avons pu mettre en avant les altérations du microbiote que présentaient les patients lupiques, avec notamment un diminution de la biodiversité, du ratio *Firmicutes*/*Bacteroidetes* ainsi que 6 taxons différentiellement abondants. Basé sur des observations empiriques, j'ai proposé une classification non-supervisée des patients lupiques, formulant l'hypothèse qu'il y avait un sous-groupe de patients avec une altération plus marquée. J'ai réalisé un *clustering* hiérarchique des patients lupiques selon la composition de leurs microbiotes et montré que parmi les deux sous-groupes de patients ainsi identifiés, un sous-groupe était composé majoritairement de patients actifs qui avaient une altération plus importante du microbiote que l'autre sous-groupe, qui était composé majoritairement de patients inactifs. Les liens entre les résultats de l'analyse non-supervisée et le statut clinique des patients (actif ou non) est un résultat particulièrement intéressant, et nous a permis de mettre en évidence une gradation de l'altération du microbiote selon l'activité de la maladie.
Une expérimentation animale a également été menée. Le but était de caractériser selon de nombreux paramètres biologiques le développement du lupus chez un modèle murin de la maladie. En raison du faible effectif (une cage de 5 souris *témoins* et une cage de 5 souris *lupiques*), les résultats concernant l'altération du microbiote en lien avec l'induction de la maladie ne sont pas facilement interprétables. Nous avons cependant trouvé des résultats concordants avec les données humaines en termes de ratio *Firmicutes*/*Bacteroidetes* et de certains taxons différentiellement abondants.
Enfin nous avons comparé les résultats que nous avions obtenus avec l'ensemble de la littérature sur le sujet, pour établir les marqueurs retrouvés de manière concordante dans plusieurs études chez l'homme et les modèles murins. Ces marqueurs ouvrent la voie à des études complémentaires pour caractériser finement l'implication du microbiote dans cette pathologie, et notamment son rôle dans l'activité du lupus et la survenue des poussées.

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 02 August 2022
DOI 10.3389/fimmu.2022.943241

OPEN ACCESS

EDITED BY
Lidan Zhao,
Peking Union Medical College Hospital (CAMS), China

REVIEWED BY
Ikhwan Rinaldi,
RSUPN Dr. Cipto Mangunkusumo, Indonesia
Hao Li,
Harvard Medical School, United States

*CORRESPONDENCE
Eya Toumi
e.toumi@alphabio.fr

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Autoimmune and Autoinflammatory Disorders, a section of the journal Frontiers in Immunology

RECEIVED 13 May 2022
ACCEPTED 04 July 2022
PUBLISHED 02 August 2022

CITATION
Toumi E, Goutorbe B, Plauzolles A, Bonnet M, Mezouar S, Militello M, Mege J-L, Chiche L and Halfon P (2022) Gut microbiota in systemic lupus erythematosus patients and lupus mouse model: a cross species comparative analysis for biomarker discovery.
Front. Immunol. 13:943241.
doi: 10.3389/fimmu.2022.943241

Gut microbiota in systemic lupus erythematosus patients and lupus mouse model: a cross species comparative analysis for biomarker discovery

Eya Toumi[1,2,3]*, Benoit Goutorbe[3,4,5], Anne Plauzolles[3], Marion Bonnet[3], Soraya Mezouar[1,2], Muriel Militello[1,2], Jean-Louis Mege[1,2,6], Laurent Chiche[7†] and Philippe Halfon[1,2,3,7†]

[1]Aix-Marseille Univ, Microbes, Evolution, Phylogénie et infection (MEPHI), Institut de recherche pour le développement (IRD), Assistance Publique-Hopitaux de Marseille (APHM), Marseille, France, [2]Institut Hospitalo-universaire (IHU)-Méditerranée Infection, Marseille, France, [3]Laboratoire Alphabio, Clinical Research and R&D Department, Marseille, France, [4]Centre de Recherche en Cancérologie de Marseille (CRCM), Aix-Marseille Univ U105, Inserm U1068, CNRS UMR7258, Institut Paoli-Calmettes, Marseille, France, [5]Université Paris-Saclay, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE), Jouy-en-Josas, France, [6]Hopital de la Conception, Immunology Department, Marseille, France, [7]Infectious and Internal Medicine Department, Hôpital Européen Marseille, Marseille, France

An increasing number of studies have provided strong evidence that gut microbiota interact with the immune system and stimulate various mechanisms involved in the pathogenesis of auto-immune diseases such as Systemic Lupus Erythematosus (SLE). Indeed, gut microbiota could be a source of diagnostic and prognostic biomarkers but also hold the promise to discover novel therapeutic strategies. Thus far, specific SLE microbial signatures have not yet been clearly identified with alteration patterns that may vary between human and animal studies. In this study, a comparative analysis of a clinically well-characterized cohort of adult patients with SLE showed reduced biodiversity, a lower *Firmicutes/Bacteroidetes* (*F/B*) ratio, and six differentially abundant taxa compared with healthy controls. An unsupervised clustering of patients with SLE patients identified a subgroup of patients with a stronger alteration of their gut microbiota. Interestingly, this clustering was strongly correlated with the disease activity assessed with the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) score ($p = 0.03$, odd ratio = 15) and the identification of specific alterations involving the *F/B* ratio and some different taxa. Then, the gut microbiota of pristane-induced lupus and control mice were analyzed for comparison with our human data. Among the six differentially abundant taxa of the human disease signature, five were common with our murine model. Finally, an exhaustive cross-species comparison between our data and previous human and murine SLE studies revealed a core-set of gut microbiome species that might constitute biomarker panels relevant for future validation studies.

KEYWORDS

systemic lupus erythematosus, gut microbiota, dysbiosis, disease activity, outcome assessment, health care, biomarkers

Introduction

Systemic lupus erythematosus (SLE) is a complex autoimmune disease characterized by a breakdown in tolerance to nuclear antigens. This leads to immune-complex deposits that cause severe inflammation in various organs such as the skin, joint, and kidney. Its broad-spectrum manifestations and its unpredictable course between active and remissive stages complicate the disease monitoring and represent a challenge to clinicians (1). SLE primarily affects women of child-bearing age, and its etiology remains unclear but there is strong evidence that genetic, hormonal, and environmental factors are involved (2). Current SLE treatments are mainly immunosuppressive drugs with unsatisfactory clinical response and functional remission rates and can lead to serious side effects (3, 4). Additionally, the long-term use of these treatments has been associated with higher incidences of more severe infections (5). There is now a crucial need to better understand the pathogenesis of SLE and propose a new therapeutic strategy without adverse effects to improve both the quality of life and survival of patients with SLE.

Recently, with the revolutionary advances in next generation sequencing (NGS) technique, emerging investigations in human and murine models have shown that disturbed microbial compositions and functions called "dysbiosis" are involved in the pathophysiology of autoimmune diseases such as inflammatory bowel disease, type 1 diabetes, rheumatoid arthritis, and multiple sclerosis (6). Growing evidence suggests that gut microbiota also play a role in SLE pathogenesis (7–9). A gut permeability called "leaky gut" was observed in lupus studies leading to altered gut barrier function (10). A decrease in beneficial bacteria such as *Bifidobacterium* (11, 12) and an increase in harmful bacteria such as *Enterococcus gallinarum* (13) and *Ruminococcus gnavus* (14), which are closely related to disease progression, were observed in both human and murine lupus. Hence, gut microbiota analysis may offer new possibilities for early diagnosis, prevention, and therapeutic approaches based on gut microbiome modulation in SLE.

However, to date, the association between gut dysbiosis and SLE activity remains unclear. Existing studies are limited to only observational case-control reports in which gut microbiome dynamics are compared to matched controls with a single time-point analysis and therefore a considerable risk of finding false positive associations. Additionally, there is discordance between human cohorts due to differences in the ethnicity and lifestyle of the populations studied. The few existing interventional studies involve only murine models that may differ in anatomy and physiology from human patients with SLE. Currently, there are no comparative studies between the two. Thus, longitudinal studies are needed to establish a common signature of gut microbiota in human and murine SLE that can serve as diagnostic and prognostic biomarkers for the disease.

Through this study, we first longitudinally investigated the dynamics of the gut microbiota of both active and inactive patients with SLE compared with a healthy population. Then, we explored the association between the gut dysbiosis and the disease activity to propose the first French gut microbiome signature of SLE. We further analyzed the murine gut microbiota in a pristane-induced lupus mouse model to identify a common and robust microbial signature of the disease between humans and mice. Finally, based on our results and those of existing studies, we propose a panel of bacterial populations commonly found to define a universal gut microbiota signature of SLE.

Materials and methods

Human study design

Stool samples from patients aged ≥18 years with a diagnosis of SLE according to the American College of Rheumatology (ACR) criteria, regardless of disease activity and ongoing treatments, were collected in the European Hospital of Marseille. Disease activity was scored based on the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) (15). Patients with SLE with severe anemia (Hb <7 g/dl) and pregnancy were excluded. From six months to one year after their first stool sample collection, some of the included patients with SLE have provided a second stool sample for the longitudinal microbiota analysis. Patients with SLE were compared with healthy controls (HCs) recruited by considering the sex-ratio of patients with SLE as well as their age range. These individuals have no known chronic pathology or any specific treatment that could disrupt their gut microbiota during the last two months preceding the stool sampling.

Animal experimental design and lupus induction

Nine-week-old female BALB/cByJ (Charles River Laboratories, L'Arbresle, Lyon, France) were housed in a controlled temperature and pressure environment. Mice were adapted to new environmental conditions for one week before the beginning of the experimental procedure. The animals were kept in cages with water and food *ad libitum*, enriched with cardboard houses with cotton squares as nests.

Animals were randomly divided into two groups, including a pristane-induced-lupus (PIL) group (n = 5) that received a single intra-peritoneal injection of 500 µl of sterile pristane oil (2, 6, 10, 14-tetramethylpentadecane, Sigma Aldrich, MO, USA) according to Satoh etal. (16) and a control (CO) group (n = 5) that received a single intra-peritoneal injection of 500 µl of sterile phosphate buffered saline (PBS, Sigma Aldrich). Blood, stool, and urine samples were collected before PBS/pristane induction

(Day 0) and at six months post-induction (M6). The animals were observed weekly for clinical monitoring. At the end of the experiment (M6), all animals were euthanized by lethal overdose of Dolethal® after an anesthetic protocol (including 90 mg/kg of ketamine® and 10 mg/kg of xylazine®).

Evaluation of SLE-like disease in PIL mice

Immunological and inflammatory analyses were performed in serum samples collected on day 0 and at M6 post-induction to validate the SLE onset. Immunological analysis included antinuclear antibody (ANA) detection determined using the indirect immunofluorescence method using commercial slides containing HEp-2 cells (Kallestad HEp-2 Cell Line Substrate, 12-well slides, Bio-Rad Laboratories, Hercules, CA) and antibodies against double-stranded-DNA (ds-DNA) quantification using the ELISA method using the mouse anti-dsDNA IgG-specific ELISA kit (Mybiosource, San Diego, CA, USA) according to the instructions of the manufacturer. For inflammatory analysis, levels of interferon (IFN)-α (PBL assay science, Piscataway, NJ, USA), tumor necrosis factor (TNF)-α (Aviva System Biology, USA) and C-reactive protein (CRP) (Aviva System Biology, USA) were measured using commercially available ELISA kits according to the instructions of the manufacturer.

A clinical assessment of arthritis was performed weekly starting two weeks after pristane induction, looking for redness and swelling in the paws. Histopathology and immunofluorescence analysis were performed in the kidneys and lungs to investigate tissue damage and immune-complex deposits.

Stool sample collection and 16S rRNA sequencing

Human and murine stool samples were collected, stored immediately in stabilizing solution (DNA/RNA shield, Zymo Research, Freiburg, Germany) and frozen at −20°C until analysis. Bacterial DNA was isolated using the ZymoBIOMICS DNA prepKit (Zymo Research) following the instructions of the manufacturer. To determine the gut microbiome composition of each sample, a metagenomic sequencing library targeting the V3-V4 regions of the 16S rRNA gene was created following Illumina's recommendations as previously described (17).

Bioinformatic processing

Sequencing reads were processed with an in-house pipeline, as previously described (17). Briefly, preprocessing and denoising were performed using Qiime2 (18) (version 2021.11) and DADA2 (19). Resulting amplicon sequence variants (ASVs) were taxonomically assigned with Kraken (20) (version 1.1) based on the NCBI RefSeq Targeted Loci database. Phylogenetic tree of ASVs were generated independently for human cohort and mice experiment was built using mafft (21) (version 7.407) and fast tree (22) (version 2.1.10) with default parameters.

Statistical analysis

Statistical analysis was performed with R (version 4.1) using the phyloseq package (version 1.36.0) (23). We ensured a minimal depth of 50 000 reads per sample, that discarded 3 of 79 available HC samples (consequently, only 76 HC samples were used for downstream analyses). We performed a rarefaction at lowest sample depth for human and murine data sets independently, resulting in 56 219 reads/sample and 62 463 reads/sample respectively. To compare microbiotas diversity and composition, we assessed alpha-diversity by Shannon index, beta-diversity by Bray-Curtis dissimilarity index which was visualized through principal component analysis (PCoA). Permanova test was performed to track the effect of clinical conditions on distances between samples. Differentially abundant taxa were identified using DESeq2 method (24) (version 1.32.0). Unsupervised classification of samples was performed based on Bray-Curtis dissimilarity indices, using hierarchical clustering with Wards linkage criteria and the two main clustered were retrieved. We then compared clustering results to disease activity and other clinical variables available (see Supplementary Data for details).

Results

Clinical and biological characteristics of SLE patients

A total of 16 SLE patients and 76 sex-age matched HCs were included. The mean age was 42 ranged from 19 to 70 years old and a female-to-male ratio of 7:1. At inclusion, SLEDAI score ranged from 0 to 12, with 9/16 patients having inactive SLE (SLEDAI=0). For the therapeutic regimen, 12/16, 4/16 and 1/16 patients received hydroxychloroquine, prednisone or immunosuppressants, respectively. Basic clinical and biological characteristics of SLE patients were shown in Table 1.

Gut microbiota of SLE patients is altered compared to HCs

To measure the similarity of gut microbial communities' composition, the beta-diversity was measured using Bray-Curtis distance on ASVs. A PCoA was used for visualizing samples projections and did not show clear distinct clustering pattern

10.3389/fimmu.2022.943241

TABLE 1 Clinical and biological characteristics of SLE patients.

ID patient	Age (years)	Sex	BMI	PGA	SLEDAI	Low complement levels	Positive anti-dsDNA titres	AHT	Type 2 diabetes	APS	Ongoing SLE treatments
001	19	M	18.7	0.12	0	no	no	no	no	no	no
002	22	F	22.8	2.1	12	yes	yes	no	no	no	HCQ CT
003	25	F	23.8	0.63	0	no	no	no	no	yes	HCQ
004	49	F	35.1	0	0	no	no	no	no	no	no
005	55	F	17.9	0.3	0	no	no	no	no	no	no
006	33	F	23.1	0	2	no	no	no	no	yes	HCQ AZA
007	70	F	23.3	0.72	2	no	yes	no	no	no	HCQ CT
008	46	F	20.7	0.75	2	no	yes	no	no	no	HCQ
009	69	F	23.1	0.27	0	no	no	no	no	no	CT
011	43	F	24.2	0.09	0	no	no	no	no	no	HCQ CT
012	33	F	21.9	0.21	4	yes	yes	no	no	no	HCQ
013	28	M	18.8	0.24	0	no	no	no	no	no	HCQ
014	55	F	19.3	1.02	0	no	no	no	no	yes	HCQ
017	35	F	19.6	0.66	0	no	no	no	no	no	HCQ
018	49	F	32	0.84	4	no	no	no	yes	no	HCQ
019	38	F	21.7	0.21	2	no	no	no	no	no	HCQ

M, male sex; F, female sex; BMI, Body mass index; PGA, Physician Global Assessment; SLEDAI, SLE Disease Activity Index; AHT, Arterial hypertension; APS, antiphospholipid syndrome; HCQ, hydroxychloroquine; CT, corticosteroids; AZA, azathioprine (immunosuppressive drug).

between SLE and HCs groups (Figure 1A). However, a permanova test revealed that the gut microbiota composition of SLE patients was significantly different from HCs ($p<0.01$). SLE patients showed a significant decrease in alpha-diversity compared to HCs regarding all qualitative, quantitative, and phylogenetic-aware metrics (Figure 1B and Figure S1). Moreover, a lower *F/B* ratio was observed in SLE patients (Figure 1C, $p<0.05$). We subsequently tracked differentially abundant taxa between SLE patients and HCs using DESeq2 to identify *de novo* biomarkers. At the phyla level, our analysis showed a significant decrease in *Tenericutes* in SLE patients ($p<0.05$). In contrast, *Tannerellaceae* family ($p<0.01$), *Alistipes* ($p<0.05$), *Flintibacter* ($p<0.05$) and *Parabacteroides* ($p<0.01$) genus were significantly abundant in SLE patients. Among *Alistipes* genus, the trend was mostly driven by one ASV that was classified as *A. onderdonkii* ($p<0.01$) and therefore this species was as well significantly more abundant in SLE patients ($p<0.001$) (Figure 1D and Figure S2).

Taken together, these findings illustrate that SLE patients have a different gut microbiota profile than HCs showing a decreased alpha-diversity and *F/B* ratio with six differentially abundant SLE biomarkers.

Unsupervised classification reveals two distinct clusters that correlates with SLE activity

We performed an unsupervised clustering based on gut microbiota compositions of SLE patients to look for subgroups of patients sharing similar gut microbiota. This analysis revealed two main clusters: Cluster 1 (referred to CL1) and Cluster 2 (referred to CL2), containing 10 and 6 SLE patients respectively. We observed that CL2 was enriched with active-SLE patients (Fisher' exact test, $p<0.05$, odd ratio=15), and was composed of patients with significantly higher SLEDAI score taken as numeric value (Wilcoxon's test, $p<0.05$). This association was not found with age, sex, BMI, treatment or enterotype excluding the possible confounding factors in the differences observed between the two clusters (Figure 2A). We measured the pairwise Bray-Curtis distances between each SLE patient, and each HC. We showed that CL2 patients were more distant to HCs than CL1 patients suggesting a dysbiosis gradient between the two clusters (Figure 2B). These alterations were subsequently observed in the *F/B* ratio, which was more disturbed in CL2 than CL1 ($p<0.05$), as well as in many differentially abundant taxa. The gut microbiota of CL2 patients was enriched with an unclassified family belonging to *Verrucomicrobia* phylum, *Desulfovibrio piger* and *Bacteroides thetaiotaomicron* species compared to CL1 patients. While some populations within the *Firmicutes* phylum were decreased including *Bacilli* class, *Clostridales* order, *Ruminococcaceae, Eubacteriaceae, Lactobacillaceae* families, *Romboutsia, Lactobacillus, Fusicatenibacter, Turicibacter* genus, *Faecalibacterium prausnitzii*, *Fusicatenibacter saccharivorans* and *Eubacterium cellulosolvens* species (Figures 2C). All these findings showed that our unsupervised analysis reveals two different clusters of patients with a marked dysbiosis gradient and correlated highly with their SLEDAI score suggesting that the gut microbiota is involved in the severity of the disease.

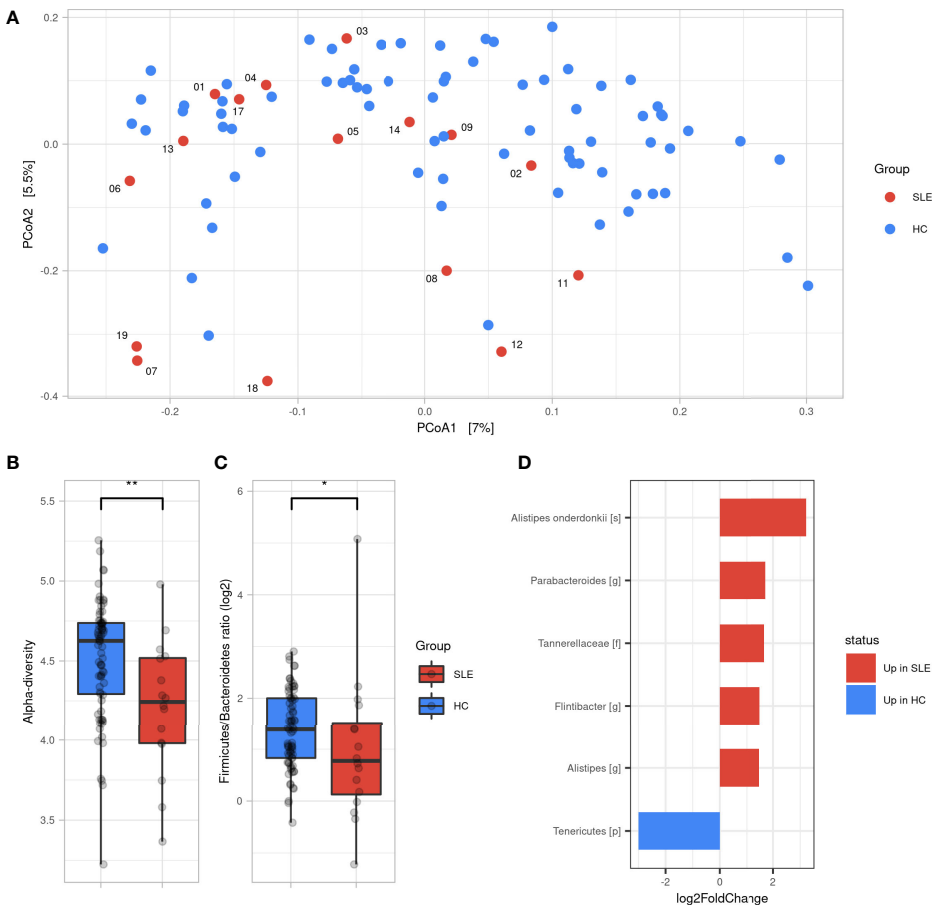


FIGURE 1
Gut microbiota difference between SLE patients and HC. **(A)** Principal coordinate analysis (PCoA) of beta-diversity based on Bray-Curtis distances. **(B)** Alpha diversity assessed by Shannon's index between SLE and HC groups. **(C)** *Firmicutes/Bacteroidetes* ratio difference between SLE and HC groups. Statistical differences between groups are shown: *p <0.05, **p <0.01 by Wilcoxon's test. **(D)** Differentially abundant taxa between SLE and HC groups identified by DESeq2: only taxa with adjusted p <0.05, absolute log2FoldChange >1 and prevalence per group >0.333 are shown. SLE, systemic lupus erythematosus; HC, healthy controls.

Gut microbiota of active SLE patients is altered compared to inactive SLE patients

Given that the gut microbiota was significantly different in SLE patients compared to HCs and that the dysbiosis was highly correlated with disease activity, we next performed a supervised gut microbiota analysis between active and inactive SLE patients. We first showed that active SLE patients were more distant to HC than inactive SLE patients and consistently observed across metrics (Figure S3). The permanova test on beta-diversity demonstrates that the composition of active SLE patients' microbiota was different than the one of inactive SLE patients ($p<0.001$) while no statistical difference was observed in alpha-diversity between groups. Active SLE patients have a significantly lower *F/B* ratio than inactive SLE group ($p<0.01$). Furthermore, as shown in Figure 2D, six differentially abundant taxa were identified including increased *Desulfovibrio piger*, *Bacteroides thetaiotaomicron* and *Ruminococcus gnavus* species and decreased *Bacilli* class, *Ruminococcaceae* and *Lactobacillaceae* families in active SLE patients compared to inactive-SLE patients. This pattern, except for *R. gnavus* species, was commonly observed in CL2 which confirms that this subgroup mainly reflects the dysbiosis that occurs in active SLE patients. Altogether, our results indicate that the gut microbiota profiling of active SLE patients were markedly different with a severe gut microbiota dysbiosis compared to inactive SLE patients.

SLE severity signature is stable over time

To ensure the robustness of the clustering, we added to the analysis the second stool sample, available from nine of our 16 SLE patients. We noted that all samples at their second time-point where highly similar to their first time point and clustered together except for one patient as shown in Figure S4A. Indeed,

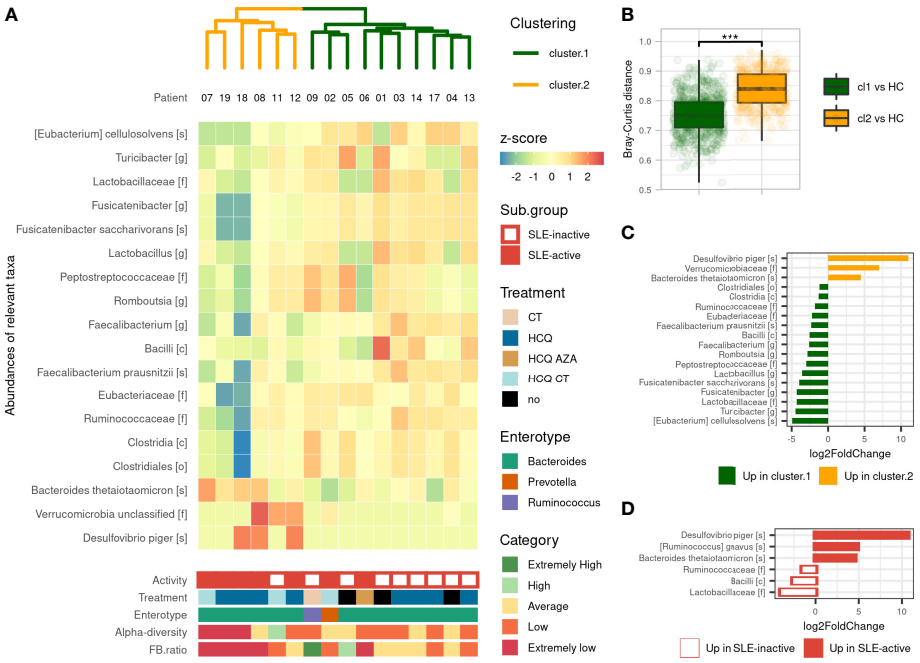


FIGURE 2
Gut microbiota's composition-based unsupervised classification of SLE patients. **(A)** Hierarchical clustering based on Bray–Curtis and Ward's linkage show two clusters of SLE patients based on their microbiota's composition and Heatmap of differentially abundant taxa illustrate differences in microbiota's composition. Activity (based on SLEDAI score), treatment and enterotype were assessed between clusters. Alpha-diversity (by Shannon's index) and *F/B* ratio were assessed and compared to HCs distribution to evaluate for comprehensive visualization. **(B)** Pairwise Bray–Curtis distances between each patient with SLE and each HC according to clustering. Statistical difference is shown. ***p <0.001 by Wilcoxon test. **(C)** Differentially abundant taxa identified by DESeq2 between the two clusters. **(D)** Differentially abundant taxa identified by DESeq2 between active and inactive SLE patients. Only taxa with adjusted p-value <0.05, absolute log2FoldChange >1 and prevalence per group >0.333 are shown. SLE, systemic lupus erythematosus; Cl1, cluster 1; Cl2, cluster 2; CT, corticosteroids; HCQ, hydroxychloroquine; AZA, azathioprine (immunosuppressive drug).

SLE patient '12' showed a dramatic change in gut microbiota composition and moved from CL2 to CL1 while no clinical changes were observed (Table S2). Taken together, our data shows that the severity of gut dysbiosis is stable over-time.

PIL mouse model has a shared gut microbiota change with SLE patients

To determine the dynamics of murine gut microbiota during lupus progression, we established a PIL mouse model presenting human SLE symptoms (See Supplementary results and Figure S5). We analyzed the microbial profiles on Day 0 (pre-diseased time-point) and at M6 post-induction (diseased-endpoint). A PCoA based on Bray Curtis distance showed that the gut microbiota of CO and PIL mice were grouped together as a single pre-diseased cluster before lupus induction. Then, at the diseased endpoint, the gut microbiota split into two clusters: a cluster regrouping the CO mice and a cluster regrouping the PIL mice ($p<0.01$), suggesting a radical change in the gut microbiota during the onset of SLE-like symptoms (Figure 3A). No significant differences in alpha-diversity were observed between groups at baseline or at the diseased endpoint. Nevertheless, the CO mouse group had higher biodiversity at the disease end point than at baseline, while this phenomenon was not observed in the PIL mouse group (Figure 3B). Similarly, we did not detect a significant difference in the *F/B* ratio between the groups at any time point (Figures 3C, D). The taxonomical analysis revealed some bacterial population alterations in the PIL mice group at the disease endpoint compared to their pre-disease-time-point and the CO mice group. We compared the murine data only with the results of our comparison analysis between patients with SLE and HCs. Our data showed that *Tenericutes* were significantly decreased in both SLE and PIL mice. Also, the *Tannerellacea* family, *Parabacteroides*, *Bacteroides*, and *Alistipes* genera were commonly increased in PIL mice and patients with SLE (Figure S6). Taken together, the gut microbiota is disrupted during lupus development in PIL mice and shares five differentially abundant biomarkers with patients with SLE.

SLE biomarkers panel proposal through existing human and murine data

To define a universal microbial biomarker of SLE, we performed an exhaustive literature review comparing the

10.3389/fimmu.2022.943241

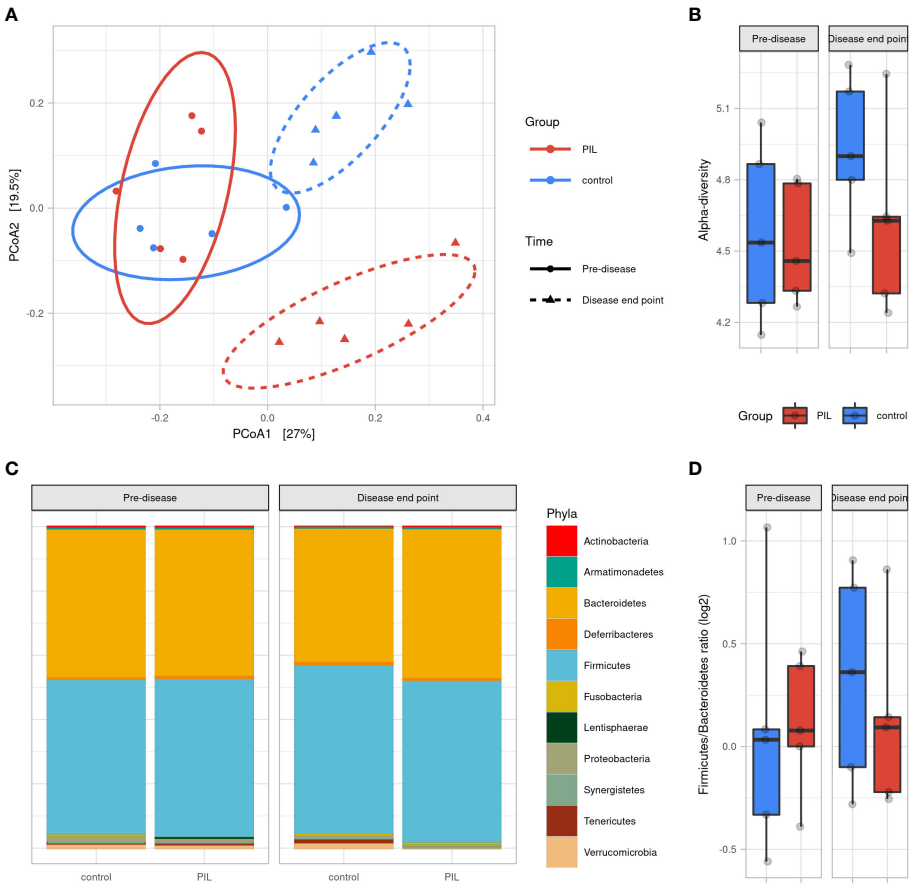


FIGURE 3
Gut microbiota difference variation overtime between PIL and Control groups. **(A)** Principal coordinate analysis (PCoA) of beta-diversitybased on Bray–Curtis distances shows that mice were uniform before induction of the disease (p = 0.6, permanova test) and strongly clustered according to groups at disease end point (6 months after induction) (p <0.01, permanova test). **(B)** Alpha diversity assessed by Shannon's index. **(C)** Gut microbiota's phyla composition according to groups and time point. **(D)** *Firmicutes/Bacteroidetes* ratio across groups and time points. PIL, pristane-induced lupus.

existing studies that have proposed an SLE gut signature compared to HCs. Among humans (n = 14) and murine studies (n = 11), 132 SLE biomarkers were identified. Biomarkers that were found in the last two studies are shown in Figure 4A. Overall, 16 biomarkers were commonly found in at least three studies, including decreased *F/B ratio* and alpha diversity as well as an increase in *Bacteroidetes*, *Proteobacteria* phyla, *Blautia*, *Bacteroides*, *Parabacteroides*, *Lactobacillus* genus, and *Ruminococcus gnavus* species and a decrease in *Firmicutes*, *Tenericutes* phyla, *Ruminococcaceae* family, *Faecalibacterium*, *Dialister*, *Bifidobacterium*, and *Desulfovibrio* genus. To track trends in our data sets that were not significant due to our relatively small sample size, we looked for the 16 most relevant biomarkers from the overall literature (raw p-values, no log2FoldChange cutoff). In our human data set, besides our signature, we found that *Bacteroidetes* and *Proteobacteria* phyla, *Bacteroides*, *Desulfovibrio* genus, and *Ruminococcus gnavus* species showed the same trend as the literature (Figure 4B). Similarly, *Bacteroides* and *Lactobacillus* genera were found coincidently in our mouse dataset (Figure 4C). Overall, we shared nine human biomarkers and seven murine biomarkers among the 16 biomarkers we commonly found.

We then established a universal panel of biomarkers including our finding according to i) the signature obtained in at least three human studies, ii) the common signature between human and mouse studies, and iii) the disease activity, as established by at least one study in our present work and two other studies (14, 25). Figure 5 shows that the *F/B* ratio and alpha diversity are the core biomarkers found in every comparison. Their decrease was reported in all studies and was associated with the disease activity. Five biomarkers were commonly found in both human and murine studies including,

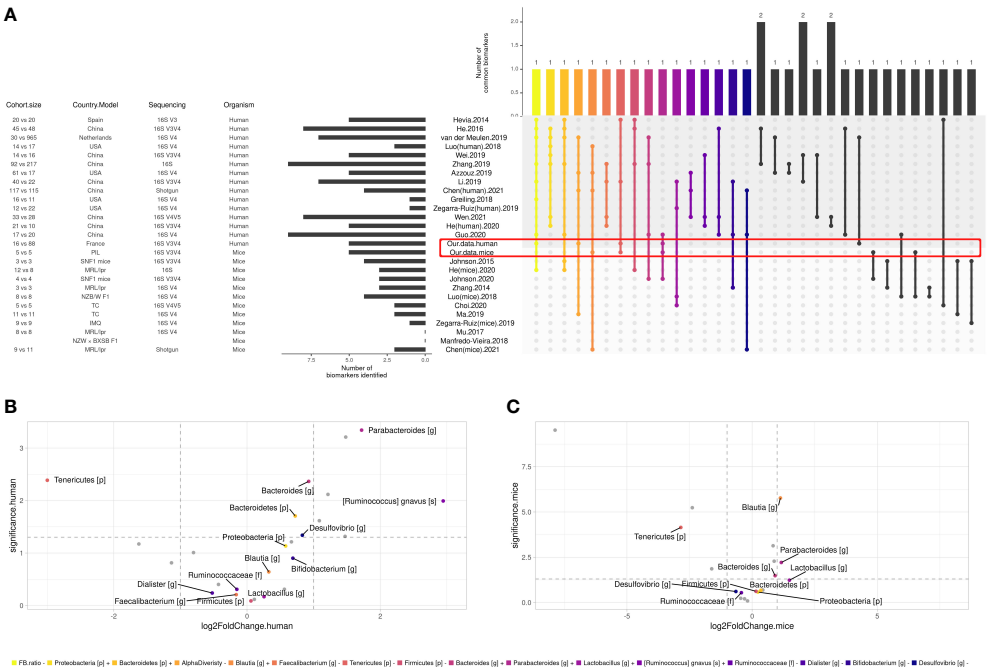


FIGURE 4
Comparison between SLE gut microbial signatures in human and mice studies across literature. **(A)** Differentially abundant taxa identified in human case-control studies and mice model. Only biomarkers identified in at least two studies are shown, and only biomarkers identified in at least three studies were attributed a color code for further investigation. **(B)** Volcano plot showing biomarkers in our human cohort and **(C)** our mice experiment, x axis shows biomarkers' log2FoldChange and y axis shows their significance (uncorrected p-value, see '*Materials and methods*').

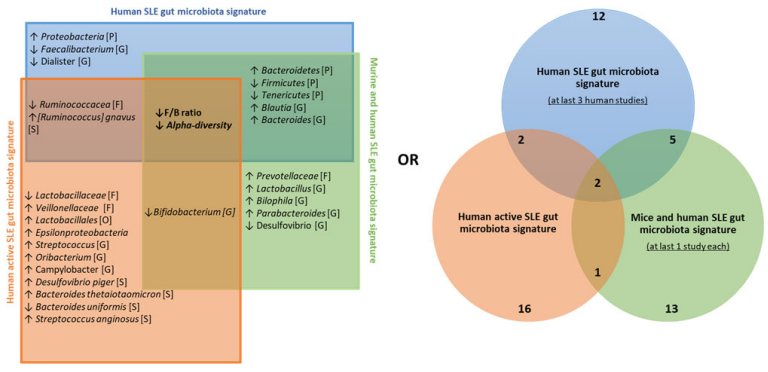


FIGURE 5
Universal panel of gut bacterial biomarkers involved in human and murine lupus and its activity. ↑, increased; ↓, decreased; [P], Phylum's taxonomy rank; [O], order's taxonomy rank [F], family's taxonomy rank; [G], Genus's taxonomy rank; [S], Species' taxonomy rank.

Bacteroidetes, *Firmicutes*, *Tenericutes* phyla, *Blautia*, and *Bacteroides* genus. Importantly, with the contribution of our data, we report a panel of 16 biomarkers related to human disease activity. As shown in Figure 5, *Ruminococcaceae* and *R. gnavus* are the most identified in human studies and are related to disease activity.

Discussion

In this study, we investigated the dynamics of gut microbiota in both human and murine lupus. We show for the first time that French patients with SLE have an imbalanced gut microbiota compared with HCs. Then, to the best of our knowledge, this

study is also the first to have performed an unsupervised approach of gut microbiota in patients with SLE, irrespective of their clinical data, to investigate the correlation between the degree of their dysbiosis and their disease activity. We show that the dysbiosis of SLE gut microbiota correlates with the SLEDAI score. Thus, we propose different gut microbial signatures of human SLE according to gut dysbiosis and disease activity. In the PIL mouse model, we show a different gut microbiota composition before and after the disease onset. We further demonstrated that some bacterial populations are commonly found in patients with SLE. Based on an exhaustive cross-species comparison between our data and previous human and murine SLE studies, we propose a core-set of gut microbiome species that might constitute biomarker panels relevant for future validation studies.

In patients with SLE, an overall decrease in alpha-diversity and a reduced *F/B* ratio were observed in SLE patients. This imbalance seems to be the main feature of SLE dysbiosis, as it has been reported by almost all previous SLE cohorts independently of ethnicity, lifestyle, or disease stage (25–28). However, a lower diversity and F/*B* ratio have been associated with several other diseases such as type 2 diabetes (29), Crohn's disease (30) or Parkinson's disease (31), indicating that these alterations are not specific to SLE but may, however, indicate a general imbalance linked to the inflammatory process of the disease. Furthermore, a range of taxa were differentially abundant in patients with SLE compared to the HC group, including a decrease in *Tenericutes*, an increase in *Alistipes Flintibacter*, *Parabacteroides* (among *Tannerellaceae* family) genus, and *Alistipes onderdonkii* species. These gut bacteria have been implicated in health and disease in several clinical and preclinical studies. Thus, the depletion of *Tenericutes* has been previously observed in two distinct studies with active and inactive SLE patients (25, 26). These bacteria have an anti-inflammatory effect and can modulate the immune system by providing gut tolerance and preventing inflammation (32). Therefore, the increasing level of *Parabacteroides* has been previously positively correlated with inflammatory cytokines involved in SLE pathogenesis such as IL-17, IL-21, IL-2R, TWEAK, IL-35, IL-10, and IFN-γ (12) suggesting that these bacteria may play a pro-inflammatory role in stimulating immune factors. Also, *Alistipes*, a relatively recent genus of the *Bacteroidetes* phylum, was found to be increased in SLE and primary Sjögren's syndrome American patients (28). *Alistipes* dysbiosis have been reported as harmful in anxiety, myalgic encephalomyelitis, chronic fatigue syndrome, depression, and colorectal cancer and beneficial in other diseases such as colitis, autism spectrum disorders and various fibrotic liver and cardiovascular disorders (33). These conflicting findings can be explained as the *Alistipes* genus consists of 13 different species that may have opposite effects. In our case, *Alistipes onderdonkii* was overabundant in the feces of patients with SLE. This strain was recently reported as a cause of abdominal infection (34) and is reported for the first time in SLE through our study. Our results are consistent with previous studies despite the difference in the cohort size and the geographical locations of patients. We provide further evidence that gut microbiota dysbiosis in SLE patients is characterized by an imbalance between beneficial and harmful bacteria.

Although the role of altered gut microbiota in SLE has been well established, no specific microbial signature in defining the degree of disease-related dysbiosis has yet been identified. Our unsupervised analysis of the gut microbiota in patients with SLE shows two main clusters. Clustering was strongly correlated with the SLEDAI score independently of age, sex, BMI, or enterotype of patients, excluding any other confounding factors. The gut bacterial composition of the CL1 sub-group was more similar to HCs compared to the CL2 sub-group, which was more distant, suggesting a gradient of dysbiosis between the two groups. The CL1 sub-group, with a minor dysbiosis, was mainly composed of inactive patients with SLE except for two patients. The first case (patient 02) was in the flare phase of the disease at inclusion and had become inactive one year later. The second case (patient 06) had a low SLEDAI score attributed only to his alopecia at inclusion, which may be related to stress or factors other than the disease. The CL2 sub-group, with a more severe dysbiosis, was composed of active patients with SLE except for one inactive patient with no data available to evaluate the disease progression. Therefore, we hypothesized that this patient may be progressing toward a flare phase, which could be preceded by a previous gut microbiota dysbiosis. Importantly, the severity of gut dysbiosis in the CL2 sub-group was mainly due to the disruption of certain bacterial populations that were not revealed in our comparison between patients with SLE and HCs. These include an increase in the *Verrucomicrobia* unclassified family, *Desulfovibrio piger*, and *Bacteroides thetaiotaomicron* species and a decrease in the *Bacilli* class, *Clostridales* order (under *Clostridia* class), *Ruminococcaceae*, *Eubacteriaceae*, *Lactobacillaceae* families, *Romboutsia*, *Lactobacillus*, *Fusicatenibacter*, *Turicibacter*, *Faecalibacterium* genus, *Faecalibacterium prausnitzii*, *Fusicatenibacter saccharivorans*, and *Eubacterium cellulosolvens* species.

D. piger has been reported as a potential gut pathobiont and have been associated with several diseases. It has been involved in the pathogenesis of inflammatory bowel disease (IBD) (35), Parkinson's disease (36) and systemic scleroderma (37). *B. thetaiotaomicron* has been previously found in patients with SLE (28) and expressed human-anti Ro60 antibodies in the blood of patients with SLE (38), which is implicated *via* molecular mimic of Ebstein–Barr virus nuclear antigen-1 in the intuition of SLE humoral auto-immunity. In parallel, the bacterial populations that were decreased in the CL2 sub-group are all part of the *Firmicutes* phylum, which may explain the lower *F/B* ratio observed. *Firmicutes* are the main producers of butyrate, which plays a central role in the generation and maintenance of Treg cells in various gut tissues. Their decrease

has been shown to be responsible for inflammatory reactions in patients with SLE. Interestingly, these bacterial populations have various beneficial roles. Among them, *F. prausnitzii* is considered one of the most important bacterial indicators of a healthy gut with anti-inflammatory effects. Its decrease has been detected in IBD, celiac disease, obesity, and diabetes (39). Similarly, the anti-inflammatory effect of *Fusicatenibacter*, particularly its *F. saccharivorans* species, has recently been demonstrated in patients and mouse models with ulcerative colitis (40) and Crohn's disease through IL-10 induction (41, 42). As well, decreasing *Romboutsia* has recently been reported as a novel microbial biomarker for early tumor generation in cancerous mucosa (43) and Crohn's disease (41). Also, the decrease of *Lactobacillus*, a probiotic strain which can modulate innate and adaptive immune responses, has also been previously reported in SLE but with conflicting results between studies (25, 44). In fact, *Lactobacillus* levels have been frequently correlated, positively or negatively, with other human chronic diseases (45). This genus includes many species that may play many different roles in disease pathogenesis that need to be further investigated in the future, particularly for SLE. Bacteria among the *Ruminococcaceae* family are producers of short-chain fatty acids (SCFAs), which are the main source of energy for colon cells (46) and protect the integrity of the intestinal epithelial cell membrane (47). Their decrease may lead to leaky gut. Recently, a meta-analysis of relevant research publications from around the world has shown a decreased abundance of *Ruminococcaceae* with SLE, especially in Chinese patients (48). Overall, our results show that a specific microbial signature in patients with SLE with more severe dysbiosis was found and correlated strongly with the SLEDAI activity score, suggesting the contribution of gut microbiota to the severity of the disease. These findings are supported by our supervised analysis of active and inactive patients with SLE. CL2 shared the microbial signature of active patients with SLE, including decreasing in *Bacilli*, *Ruminococcaceae*, *Lactobacillaceae*, and increasing in *D. piger*, *B. thetaiotaomicron*, suggesting that these populations are strongly associated with the severity of the disease. While *R. gnavus* was only significantly increased in active patients with SLE. Azzouz et al. have shown that the intestinal expansion of this bacteria reflects the extent of the disease activity in lupus nephritis patients (14).

Then, we investigated a potential common microbial dysbiosis signature between human and murine SLE. The PIL mouse model is characterized by typical ANA, clinical manifestations, and organ involvement similar to human SLE characteristics (49, 50), making it a relevant model for studying gut microbiota. We are the first report a gut microbiota signature in the PIL mouse model. Metagenomic data from the PIL mouse model were analyzed and found to support our findings in patients with SLE despite the small number of mice. Interestingly, five biomarkers among six were shared between patients with SLE and the PIL mouse model, including *Tenericutes*, *Tannerellaceae*, *Parabacteroides*, *Bacteroides*, and *Alistipes*. To date, only two studies have investigated both human and murine gut microbiota in SLE. Luo et al. have identified that only *Lachnospiraceae* was commonly found extended in both MRL/lpr mice and American patients with SLE (51). Then, greater consensus was commonly found between MRL/lpr and Chinese patients with SLE, including 17 species (52). In the same study, more signatures in pathway analysis were shared, including pathways of L-arginine, L-ornithine, tryptophan, and menaquinol biosynthesis that were related to SLE. More consensus is still needed between humans and mice with a larger number of patients with SLE and a mouse model to be able to continue using mice to model human disease in interventional investigations.

Our study is not without limitations. In our human cohort, the active patients with SLE have mostly mild and moderate activity. We minimized this bias through our unsupervised analysis, which was able to define patients with SLE according to their dysbiosis. It seems important to note that the exact composition of clusters may differ if samples are added or removed from the dataset and is sensitive to methodological choices, notably to distance metric and clustering linkage strategy. Also, patients with SLE were enrolled while already being diagnosed and treated. Thus, we did not investigate the impact of treatment on SLE-associated dysbiosis because we did not have their stool samples before the beginning of treatments, and the small size of our cohort precludes any definitive conclusions and warrants further studies. In our murine study, because of the unavailability of disease activity score in mice due to the small number of animals, we only compared these data to those of patients with SLE in comparison to HCs. It should also be noted that we cannot rule out the possibility that cage effects are behind the differences observed between the two groups of mice and that we do not have additional cages for each group. Longitudinal investigations in a larger number of PIL mice distributed in different cages are needed in order to establish an association between changes in the gut microbiota and the establishment of SLE at different time points. Indeed, the mechanistic link between disease susceptibility and gut microbiota changes needs to be explored in this model. Also, it remains very uncertain whether gut microbiota dysbiosis is either a causative factor or a consequence of SLE disease or both. Therefore, the identification of specific bacteria responsible for the dysbiotic state in SLE may provide a better insight into the underlying mechanism. For that proposal, several thoughtful approaches can be considered, including colonization of germ-free mice with gut bacterial populations associated with the disease, as proposed by our panel, which might offer more insight into the role of these bacteria in the disease pathogenesis.

Despite our findings, which are consistent with some existing studies, the current literature regarding a common signature of gut microbiota dysbiosis in SLE is at present ambiguous. This may be due to the lack of comparative

studies between humans and mice, as discussed above, and according to the disease activity. We propose a representative pattern of gut microbiota biomarkers, which illustrates biomarker panels that are commonly found according to existing studies. Importantly, *Ruminococcaceae*, *Bifidobacterium*, and *R. gnavus* seem to play a crucial role in the severity of SLE and should be the target of future investigations to better understand the mechanisms involved. These bacterial populations are possibly trigger an auto-immune response by molecular mimicry or by influencing the Th17/Treg balance, resulting in regulatory and/or effector responses in SLE. It is a common phenomenon that a leaky bacterial product or bacteria translocation, characteristic of increased permeability of the gut, primes or educates the immune system not only in the gut but also in the entire body. Functional validation assays are needed to demonstrate the mechanistic approaches of the bacterial populations proposed in our panel and need to be enriched by other larger comparative studies.

Data availability statement

The datasets presented in this study can be found in online repositories. The name of the repository and accession number can be found below: EBI European Nucleotide Archive; PRJEB52971.

Ethics statement

The studies involving human participants were reviewed and approved by ANSM and CPP Nord-Ouest IV. The patients/participants provided their written informed consent to participate in this study. The animal study was reviewed and approved by the animal experimentation ethics committee under reference number APAFIS #26184.

Author contributions

Conceptualization: LC, AP, and ET. DNA extraction and samples sequencing: ET and MB. Animal experimentation: ET and MM. Bioinformatic and statistical analyses: BG. Data interpretation: ET and BG. Writing and original manuscript preparation: ET and BG. Review and editing: LC, AP, MB, SM, and MM. Final manuscript validation: LC, AP, ET, J-LM and PH. Supervision: AP, J-LM, LC, and PH. Funding acquisition: LC and PH. All authors discussed the results and commented on the manuscript. All authors contributed to the article and approved the submission version. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

The study was funded by the THELLIE crowfunding plateform (https://thellie.org/lupuslivinglab) and funds from European Hospital Marseille.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The funders, private donors, patient association, and pharmaceutical laboratory had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.943241/full#supplementary-material

References

1. Thong B, Olsen NJ. Systemic lupus erythematosus diagnosis and management. *Rheumatology* (2017) 56:i3–13. doi: 10.1093/rheumatology/kew401

2. Tsokos GC. Autoimmunity and organ damage in systemic lupus erythematosus. *Nat Immunol* (2020) 21:605–14. doi: 10.1038/s41590-020-0677-6

3. Nasonov E, Soloviev S, Davidson JE, Lila A, Togizbayev G, Ivanova R, et al. Standard medical care of patients with systemic lupus erythematosus (SLE) in large specialised centres: data from the Russian federation, Ukraine and republic of Kazakhstan (ESSENCE). *Lupus Sci Med* (2015) 2:e000060. doi: 10.1136/lupus-2014-000060

4. Kiriakidou M, Ching CL. Systemic lupus erythematosus. *Ann Intern Med* (2020) 172:ITC81–96. doi: 10.7326/AITC202006020

5. Tektonidou MG, Wang Z, Dasgupta A, Ward MM. Burden of serious infections in adults with systemic lupus erythematosus: A national population-based study, 1996–2011. *Arthritis Care Res* (2015) 67:1078–85. doi: 10.1002/acr.22575

6. Rosser EC, Mauri C. A clinical update on the significance of the gut microbiota in systemic autoimmunity. *J Autoimmun* (2016) 74:85–93. doi: 10.1016/j.jaut.2016.06.009

7. Colucci R, Moretti S. Implication of human bacterial gut microbiota on immune-mediated and autoimmune dermatological diseases and their comorbidities: A narrative review. *Dermatol Ther* (2021) 11:363–84. doi: 10.1007/s13555-021-00485-0

8. Li R, Meng X, Chen B, Zhao L, Zhang X. Gut microbiota in lupus: a butterfly effect? *Curr Rheumatol Rep* (2021) 23:27. doi: 10.1007/s11926-021-00986-z

9. Zhang S-X, Wang J, Chen J-W, Zhang M-X, Zhang Y-F, Hu F-Y, et al. The level of peripheral regulatory T cells is linked to changes in gut commensal microflora in patients with systemic lupus erythematosus. *Ann Rheumatol Dis* (2021) 80:e177–7. doi: 10.1136/annrheumdis-2019-216504

10. Battaglia M, Garrett-Sinha LA. Bacterial infections in lupus: Roles in promoting immune activation and in pathogenesis of the disease. *J Transl Autoimmun* (2021) 4:100078. doi: 10.1016/j.jtauto.2020.100078

11. Zhang H, Liao X, Sparks JB, Luo XM. Dynamics of gut microbiota in autoimmune lupus. *Appl Environ Microbiol* (2014) 80:7551–60. doi: 10.1128/AEM.02676-14

12. Guo M, Wang H, Xu S, Zhuang Y, An J, Su C, et al. Alteration in gut microbiota is associated with dysregulation of cytokines and glucocorticoid therapy in systemic lupus erythematosus. *Gut Microbes* (2020) 11:1758–73. doi: 10.1080/19490976.2020.1768644

13. Vieira SM, Hiltensperger M, Kumar V, Zegarra-Ruiz D, Dehner C, Khan N, et al. Translocation of a gut pathobiont drives autoimmunity in mice and humans. *Science* (2018) 359:1156–61. doi: 10.1126/science.aar7201

14. Azzouz D, Omarbekova A, Heguy A, Schwudke D, Gisch N, Rovin BH, et al. Lupus nephritis is linked to disease-activity associated expansions and immunity to a gut commensal. *Ann Rheumatol Dis* (2019) 78:947–56. doi: 10.1136/annrheumdis-2018-214856

15. Romero-Diaz J, Isenberg D, Ramsey-Goldman R. Measures of adult systemic lupus erythematosus. *Arthritis Care Res* (2011) 63. doi: 10.1002/acr.20572

16. Satoh M, Kumar A, Kanwar YS, Reeves WH. Anti-nuclear antibody production and immune-complex glomerulonephritis in BALB/c mice treated with pristane. *Proc Natl Acad Sci* (1995) 92:10934–8. doi: 10.1073/pnas.92.24.10934

17. Plauzolles A, Toumi E, Bonnet M, Pénaranda G, Bidaut G, Chiche L, et al. Human stool preservation impacts taxonomic profiles in 16S metagenomics studies. *Front Cell Infect Microbiol* (2022) 12:722886. doi: 10.3389/fcimb.2022.722886

18. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* (2019) 37:852–7. doi: 10.1038/s41587-019-0209-9

19. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from illumina amplicon data. *Nat Methods* (2016) 13:581–3. doi: 10.1038/nmeth.3869

20. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* (2014) 15:R46. doi: 10.1186/gb-2014-15-3-r46

21. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* (2002) 30:3059–66. doi: 10.1093/nar/gkf436

22. Price CJ. The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann N Y Acad Sci* (2010) 1191:62–88. doi: 10.1111/j.1749-6632.2010.05444.x

23. McMurdie PJ, Holmes S. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PloS One* (2013) 8:e61217. doi: 10.1371/journal.pone.0061217

24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* (2014) 15:550. doi: 10.1186/s13059-014-0550-8

25. Li Y, Wang H-F, Li X, Li H-X, Zhang Q, Zhou H-W, et al. Disordered intestinal microbes are associated with the activity of systemic lupus erythematosus. *Clin Sci* (2019) 133:821–38. doi: 10.1042/CS20180841

26. Hevia A, Milani C, López P, Cuervo A, Arboleya S, Duranti S, et al. Intestinal dysbiosis associated with systemic lupus erythematosus. *mBio* (2014) 5:e01548–01514. doi: 10.1128/mBio.01548-14

27. He Z, Shao T, Li H, Xie Z, Wen C. Alterations of the gut microbiome in Chinese patients with systemic lupus erythematosus. *Gut Pathog* (2016) 8:64. doi: 10.1186/s13099-016-0146-9

28. van der Meulen TA, Harmsen HJM, Vila AV, Kurilshikov A, Liefers SC, Zhernakova A, et al. Shared gut, but distinct oral microbiota composition in primary sjögren's syndrome and systemic lupus erythematosus. *J Autoimmun* (2019) 97:77–87. doi: 10.1016/j.jaut.2018.10.009

29. Larsen N, Vogensen FK, van denBerg FWJ, Nielsen DS, Andreasen AS, Pedersen BK, et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS One* (2010) 5:e9085. doi: 10.1371/journal.pone.0009085

30. Man SM, Kaakoush NO, Mitchell HM. The role of bacteria and pattern-recognition receptors in crohn's disease. *Nat Rev Gastroenterol Hepatol* (2011) 8:152–68. doi: 10.1038/nrgastro.2011.3

31. Romano S, Savva GM, Bedarf JR, Charles IG, Hildebrand F, Narbad A. Meta-analysis of the parkinson's disease gut microbiome suggests alterations linked to intestinal inflammation. *NPJ Park Dis* (2021) 7:27. doi: 10.1038/s41531-021-00156-z

32. Vieira JRP, Rezende AT, de O, Fernandes MR, da SilvaNA. Intestinal microbiota and active systemic lupus erythematosus: a systematic review. *Adv Rheumatol* (2021) 61:42. doi: 10.1186/s42358-021-00201-8

33. Parker BJ, Wearsch PA, Veloo ACM, Rodriguez-Palacios A. The genus alistipes: Gut bacteria with emerging implications to inflammation, cancer, and mental health. *Front Immunol* (2020) 11:906. doi: 10.3389/fimmu.2020.00906

34. Cobo F, Foronda C, Pérez-Carrasco V, Martin-Hita L, García-Salcedo JA, Navarro-Marí JM. First description of abdominal infection due to alistipes onderdonkii. *Anaerobe* (2020) 66:102283. doi: 10.1016/j.anaerobe.2020.102283

35. Jia W, Whitehead RN, Griffiths L, Dawson C, Bai H, Waring RH, et al. Diversity and distribution of sulphate-reducing bacteria in human faeces from healthy subjects and patients with inflammatory bowel disease. *FEMS Immunol Med Microbiol* (2012) 65:55–68. doi: 10.1111/j.1574-695X.2012.00935.x

36. Murros KE, Huynh VA, Takala TM, Saris PEJ. Desulfovibrio bacteria are associated with parkinson's disease. *Front Cell Infect Microbiol* (2021) 11:652617. doi: 10.3389/fcimb.2021.652617

37. Bellocchi C, Fernández-Ochoa Á, Montanelli G, Vigone B, Santaniello A, Milani C, et al. Microbial and metabolic multi-omic correlations in systemic sclerosis patients. *Ann N Y Acad Sci* (2018) 1421:97–109. doi: 10.1111/nyas.13736

38. Greiling TM, Dehner C, Chen X, Hughes K, Iñiguez AJ, Boccitto M, et al. Commensal orthologs of the human autoantigen Ro60 as triggers of autoimmunity in lupus. *Sci Transl Med* (2018) 10:eaan2306. doi: 10.1126/scitranslmed.aan2306

39. Miquel S, Martín R, Rossi O, Bermúdez-Humarán L, Chatel J, Sokol H, et al. Faecalibacterium prausnitzii and human intestinal health. *Curr Opin Microbiol* (2013) 16:255–61. doi: 10.1016/j.mib.2013.06.003

40. Takeshita K, Mizuno S, Mikami Y, Sujino T, Saigusa K, Matsuoka K, et al. A single species of clostridium subcluster XIVa decreased in ulcerative colitis patients. *Inflamm Bowel Dis* (2016) 22:2802–10. doi: 10.1097/MIB.0000000000000972

41. Qiu X, Zhao X, Cui X, Mao X, Tang N, Jiao C, et al. Characterization of fungal and bacterial dysbiosis in young adult Chinese patients with crohn's disease. *Ther Adv Gastroenterol* (2020) 13:1756284820971202. doi: 10.1177/1756284820971202

42. Gryaznova MV, Solodskikh SA, Panevina AV, Syromyatnikov MY, Dvoretskaya Y, Sviridova TN, et al. Study of microbiome changes in patients with ulcerative colitis in the central European part of Russia. *Heliyon* (2021) 7:e06432. doi: 10.1016/j.heliyon.2021.e06432

43. Mangifesta M, Mancabelli L, Milani C, Gaiani F, de'Angelis N, de'Angelis GL, et al. Mucosal microbiota of intestinal polyps reveals putative biomarkers of colorectal cancer. *Sci Rep* (2018) 8:13974. doi: 10.1038/s41598-018-32413-2

44. Mu Q, Zhang H, Liao X, Lin K, Liu H, Edwards MR, et al. Control of lupus nephritis by changes of gut microbiota. *Microbiome* (2017) 5:73. doi: 10.1186/s40168-017-0300-8

45. Heeney DD, Gareau MG, Marco ML. Intestinal lactobacillus in health and disease, a driver or just along for the ride? *Curr Opin Biotechnol* (2018) 49:140–7. doi: 10.1016/j.copbio.2017.08.004

46. Canfora EE, Jocken JW, Blaak EE. Short-chain fatty acids in control of body weight and insulin sensitivity. *Nat Rev Endocrinol* (2015) 11:577–91. doi: 10.1038/nrendo.2015.128

47. Peng L, Li Z-R, Green RS, Holzman IR, Lin J. Butyrate enhances the intestinal barrier by facilitating tight junction assembly *via* activation of AMP-activated protein kinase in caco-2 cell monolayers. *J Nutr* (2009) 139:1619–25. doi: 10.3945/jn.109.104638

48. Xiang S, Qu Y, Qian S, Wang R, Wang Y, Jin Y, et al. Association between systemic lupus erythematosus and disruption of gut microbiota: a meta-analysis. *Lupus Sci Med* (2022) 9:e000599. doi: 10.1136/lupus-2021-000599

49. Calvani N, Caricchio R, Tucci M, Sobel ES, Silvestris F, Tartaglia P, et al. Induction of apoptosis by the hydrocarbon oil pristane: Implications for pristane-induced lupus. *J Immunol* (2005) 175:4777–82. doi: 10.4049/jimmunol.175.7.4777

50. Freitas EC, de Oliveira MS, Monticielo OA. Pristane-induced lupus: considerations on this experimental model. *Clin Rheumatol* (2017) 36:2403–14. doi: 10.1007/s10067-017-3811-6

51. Luo XM, Edwards MR, Mu Q, Yu Y, Vieson MD, Reilly CM, et al. Gut microbiota in human systemic lupus erythematosus and a mouse model of lupus. *Appl Environ Microbiol* (2018) 84:e02288–17. doi: 10.1128/AEM.02288-17

52. Chen B, Jia X-M, Xu J-Y, Zhao L-D, Ji J-Y, Wu B-X, et al. An autoimmunogenic and proinflammatory profile defined by the gut microbiota of patients with untreated systemic lupus erythematosus. *Arthritis Rheumatol Hoboken NJ* (2021) 73:232–43. doi: 10.1002/art.41511

5 - CONCLUSIONS

Dans cette thèse, nous avons commencé par introduire les enjeux de l'analyse du microbiote humain dans un contexte clinique, ainsi que les solutions qu'apporte le séquençage à haut débit pour y répondre. Nous avons cependant constaté qu'il pourrait être intéressant d'utiliser des alternatives aux deux stratégies de séquençage dominantes, qui seraient à la fois plus résolutives que le métabarcoding, moins chères que la métagénomique *shotgun*, et insensible aux contaminations par l'ADN de l'hôte (pour le microbiote vaginal notamment). J'ai ainsi mené deux projets distincts pour répondre à cette problématique : le premier a permis d'évaluer la métagénomique *shotgun* à faible profondeur (SSM pour *shallow shotgun metagenomics*) pour l'analyse du microbiote intestinal, et le deuxième a exploré une approche de métabarcoding multi-marqueurs pour caractériser le microbiote vaginal.

Dans le chapitre 2, nous avons d'abord montré la nécessité de filtrer efficacement les résultats bruts de l'alignement pour construire des profils taxonomiques fiables, et proposé une méthode utilisant un apprentissage automatique largement plus performante que les méthodes classiques. Nous avons ensuite regardé l'impact de la profondeur de séquençage sur les profils taxonomiques obtenus et la distinction entre groupes de patients, et conclu que la perte d'information liée à la faible profondeur concernait principalement les taxa rares, que les structures d'α et de β-diversités étaient conservées, et que les classifications de patients étaient possibles en SSM, à moins que celles-ci ne reposent sur des populations sous-abondantes. Nous avons ainsi conclu que la SSM était adaptée pour l'analyse de la composition du microbiote intestinal humain, constituant une alternative intéressante pour la recherche clinique, car plus résolutive que le métabarcoding et moins chère que la métagénomique *shotgun*, ce qui peut permettre d'analyser un plus grand nombre d'échantillons.

Dans le chapitre 3, nous avons mis en évidence la complémentarité de différents marqueurs utilisés en métabarcoding, et développé plusieurs méthodes pour intégrer les profils taxonomiques provenant des différents marqueurs pour construire un profil taxonomique consensus tirant parti du pouvoir discriminant et de l'universalité de chaque marqueur. Nous avons montré l'intérêt de cette approche dans le cadre du microbiote vaginal, qui permet de produire des profils plus proches de ceux attendus, et de mieux discriminer les différentes classes du microbiote vaginal.

Enfin nous avons présenté dans le chapitre 4 l'analyse de deux jeux de données d'intérêt biomédical. Dans le premier article, nous avons comparé et classé différentes méthodes pour la conservation des échantillons selon leur performance, et établi un lien entre certains traits phénotypiques des genres bactériens et l'altération de leurs abondances relatives pour expliquer les biais liés aux méthodes de conservation des échantillons. Dans le deuxième article, nous avons déterminé une signature microbiotique des patients atteints de lupus érythémateux systémique, et montré que les patients actifs avaient une dysbiose plus marquée. Nous avons ensuite considéré la signature obtenue chez l'homme, celle obtenue chez un modèle murin de la maladie, ainsi que les résultats de l'ensemble de la littérature sur le sujet, pour déterminer un panel de bactéries potentiellement impliquées dans le développement de cette pathologie qui pourra servir de socle pour des expérimentations futures.

Les deux stratégies de séquençage étudiées au cours de mon travail représentent des alternatives plausibles et pertinentes pour les écosystèmes d'intérêt abordés dans cette thèse, à savoir le microbiote intestinal pour la SSM et le microbiote vaginal pour le métabarcoding multi-marqueurs.

Concernant le microbiote intestinal, la métagénomique *shotgun* est souvent plébiscitée par

rapport au métabarcoding, notamment parce qu'elle offre une meilleure résolution taxonomique et la possibilité de réaliser des analyses fonctionnelles. Cependant, l'écart de coûts entre les deux stratégies incite de nombreux chercheurs à utiliser le métabarcoding. La SSM, dont les coûts de séquençage se rapprochent de ceux du métabarcoding, constitue une alternative qui devrait faciliter le développement de l'approche *shotgun*. Le métabarcoding multi-marqueurs pourrait également s'appliquer à l'analyse du microbiote intestinal, dans l'optique de combler les lacunes du métabarcoding ciblant uniquement le gène 16S, qui est largement utilisé. Comparativement au métabarcoding ciblant le gène 16S, l'approche multi-marqueurs pourrait améliorer la résolution taxonomique et l'estimation des abondances relatives. Un exemple concret serait de permettre une meilleure résolution taxonomique au sein de l'ordre des *Enterobacterales*, qui contient de nombreux pathogènes, et dont le gène 16S est très conservé. Cependant, comparativement à la SSM, le métabarcoding multi-marqueurs n'ouvre pas la voie de l'analyse fonctionnelle.
Pour le microbiote vaginal, comme pour les microbiotes cutané et salivaire, le fort taux de contamination par l'ADN de l'hôte dans les prélèvements apporte une contrainte supplémentaire qui rend la métagénomique *shotgun* peu rentable. Il serait nécessaire de séquencer un très grand nombre de *reads* pour espérer obtenir suffisamment de *reads* bactériens utiles à l'analyse du microbiote. S'il existe des méthodes biochimiques pour réduire le taux d'ADN humain, celles-ci biaisent de manière non-négligeable les abondances relatives des espèces [1, 102]. En revanche, ces écosystèmes étant moins diversifiés que le microbiote intestinal, le nombre de *reads* bactériens nécessaires pour construire les profils taxonomiques est plus faible.
Les stratégies de séquençage présentées peuvent être pertinentes pour l'analyse des écosystèmes microbiens autres que le microbiote humain. La SSM est principalement limitée par la nécessité de disposer d'un catalogue de référence exhaustif sur lequel aligner les données. Dès lors qu'un tel catalogue est disponible, les problématiques abordées dans le chapitre 2 peuvent s'appliquer à tout type d'écosystèmes microbiens. Ce type de catalogues est aujourd'hui disponible ou en cours de construction pour de nombreux types d'écosystèmes, en lien par exemple avec l'agronomie [190, 191, 48], permettant ainsi le développement de cette approche. Comme pour le microbiote intestinal humain, la profondeur de séquençage nécessaire dépendra de la problématique biologique, de la complexité de l'écosystème et du niveau de détail attendu concernant les taxa rares.
Le métabarcoding multi-marqueur pourrait être facilement applicable, et s'avérer utile dans un grand nombre de contextes biologiques. Les méthodes multi-marqueurs peuvent permettre de tirer parti de l'universalité, du pouvoir discriminant et de la complétude de bases de données de chaque marqueur, ainsi que de niveler les biais liés à leur nombre de copies. Que ce soit sur les écosystèmes microbiens procaryotes [137, 125, 84, 123, 110], ou eucaryotes [168, 174, 35], la complémentarité de différents marqueurs a été mise en évidence dans plusieurs contextes, qui sont autant d'exemples pour lesquels les méthodes que nous avons développées seraient pertinentes.

Les technologies de séquençage évoluent constamment. Nous avons dans cette thèse considéré uniquement les technologies de séquençage à haut débit de type Illumina. Cette technologie de séquençage, caractérisée par des *reads* courts et contenant peu d'erreurs de séquençage, est aujourd'hui la technologie la plus utilisée. L'apparition de concurrents, qui produisent des données similaires à celles d'Illumina [46, 10] est susceptible de faire baisser les coûts de séquençage. Le rapport entre les coûts de séquençage des différentes stratégies restera le même

mais pourrait, à terme, être moins déterminant dans les choix méthodologiques. Par ailleurs, le séquençage de *reads* longs, permis par les technologies de PacBio et d'Oxford Nanopore, offrent d'autres possibilités. En métabarcoding, ces technologies rendent notamment possible de lire des séquences 16S de pleine longueur (~ 1500bp) au lieu de se limiter à une portion du gène. Ceci permet une meilleure résolution taxonomique, mais ne résout pas les problèmes d'universalité des amorces et du biais du nombre de copies (mentionnés dans la table 1.1, page 19). Ces technologies sont également intéressantes en métagénomique *shotgun*, particulièrement dans l'optique de construire des catalogues de gènes et génomes. La longueur des *reads* rend l'étape d'assemblage beaucoup plus facile, notamment pour les régions répétées et conservées[192]. Ainsi, ces technologies vont permettre dans les prochaines années d'améliorer considérablement les catalogues de gènes et génomes, notamment concernant les organismes non cultivables. L'amélioration de ces catalogues bénéficiera aux analyses métagénomiques, et en particulier à la SSM.

Les travaux présentés dans cette thèse ont permis de montrer la pertinence de stratégies de séquençage intermédiaires pour l'analyse des microbiotes humains, à la fois peu coûteuses et suffisamment résolutives. Leur faible coût en font des méthodes adaptées pour l'analyse des microbiotes humains à grande échelle, facilitant tant la recherche clinique que la mise en application dans des parcours de soin. Ces méthodes produisent des profils taxonomiques précis et fiables, pouvant servir pour discriminer les patients dans certaines situations cliniques d'intérêt. Les choix méthodologiques, tant pour la stratégie de séquençage que l'analyse bioinformatique, doivent être guidés par les problématiques cliniques. Ce travail s'inscrit dans l'ensemble des efforts nécessaires pour la mise en application des connaissances grandissantes sur le microbiote humain, au service de la médecine.

Communications scientifiques

Articles :

- **Goutorbe B**, Abraham A-L, Mariadassou M, Plauzolles A, Bidaut G, Halfon P, Schbath S. Characterizing the limits of shallow shotgun metagenomics for taxonomic profiling of human gut microbiota in clinical studies. Article under review. Preprint available at doi : 10.21203/rs.3.rs-1306026.
- Toumi E, **Goutorbe B**, Plauzolles A, Bonnet M, Mezouar S, Militello M, Mège J-L, Chiche L[†], Halfon P[†], Gut microbiota in systemic lupus erythematosus patients and lupus mouse model : A cross species comparative analysis for biomarker discovery. Front Cell Immunol. Article accepted on july 4thth 2022 and available soon.
- Plauzolles A[†], Toumi E[†], Bonnet M, Pénaranda G, Bidaut G, Chiche L, Allardet-Servent J, Retornaz F, **Goutorbe B**[‡], Halfon P[‡]. Human Stool Preservation Impacts Taxonomic Profiles in 16S Metagenomics Studies. Front Cell Infect Microbiol. 2022 Feb 8 ;12 :722886. doi : 10.3389/fcimb.2022.722886. PMID : 35211421 ; PMCID : PMC8860989.

Présentation orale :

- **Goutorbe B**, Abraham A-L, Mariadassou M, Plauzolles A, Bidaut G, Halfon P, Schbath S. Shallow Shotgun Metagenomics as a cost-effective and accurate alternative to WGS for taxonomic profiling and clinical diagnosis. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), congrès en ligne, du 6 au 9 juillet 2021.

Posters :

- **Goutorbe B**, Abraham A-L, Mariadassou M, Bonnet M, Plauzolles A, Bidaut G, Halfon P, Schbath S. Combining marker genes in amplicon sequencing : a novel approach enhancing taxonomic profiling of microbial ecosystems. 10th Congress of the International Symbiosis Society & 3rd International Conference on Holobionts, à Lyon du 25 au 29 juillet 2022.
- **Goutorbe B**, Abraham A-L, Mariadassou M, Loux V, Rué O, Plauzolles A, Bidaut G, Halfon P, Schbath S. Shallow shotgun metagenomics to study gut microbiota : contributions of machine learning. Colloque "Sciences numériques et Intelligence Artificielle pour la Santé à Aix-Marseille Université", à Marseille, les 25 et 26 novembre 2021.
- **Goutorbe B**, Abraham A-L, Mariadassou M, Loux V, Rué O, Plauzolles A, Bidaut G, Halfon P, Schbath S. Shallow sequencing : a cost-effective and accurate alternative to WGS for taxonomic profiling ? Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), congrès en ligne, du 30 juin au 3 juillet 2020.

Bibliographie

[1] S. Ahannach, L. Delanghe, I. Spacova, S. Wittouck, W. Van Beeck, I. De Boeck, and S. Lebeer. Microbial enrichment and storage for metagenomics of vaginal, skin, and saliva samples. *iScience*, 24(11) :103306, Oct. 2021.

[2] J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society : Series B (Methodological)*, 44(2) :139–160, 1982. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1982.tb01195.x.

[3] J. L. Alexander, I. D. Wilson, J. Teare, J. R. Marchesi, J. K. Nicholson, and J. M. Kinross. Gut microbiota modulation of chemotherapy efficacy and toxicity. *Nature Reviews. Gastroenterology & Hepatology*, 14(6) :356–365, June 2017.

[4] A. Almeida, A. L. Mitchell, M. Boland, S. C. Forster, G. B. Gloor, A. Tarkowska, T. D. Lawley, and R. D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753) :499–504, Apr. 2019.

[5] A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, and R. D. Finn. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39(1) :105–114, Jan. 2021.

[6] M. Almeida, M. Pop, E. Le Chatelier, E. Prifti, N. Pons, A. Ghozlane, and S. D. Ehrlich. Capturing the most wanted taxa through cross-sample correlations. *The ISME Journal*, 10(10) :2459–2467, Oct. 2016. Number : 10 Publisher : Nature Publishing Group.

[7] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11) :1144–1146, Nov. 2014. Number : 11 Publisher : Nature Publishing Group.

[8] F. E. Angly, P. G. Dennis, A. Skarshewski, I. Vanwonterghem, P. Hugenholtz, and G. W. Tyson. CopyRighter : a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, 2(1) :11, Dec. 2014.

[9] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson. Grinder : a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, 40(12) :e94–e94, July 2012.

[10] S. Anslan, V. Mikryukov, K. Armolaitis, J. Ankuda, D. Lazdina, K. Makovskis, L. Vesterdal, I. K. Schmidt, and L. Tedersoo. Highly comparable metabarcoding results from MGI-Tech and Illumina sequencing platforms. *PeerJ*, 9 :e12254, 2021.

[11] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, MetaHIT Consortium, M. Antolín, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux,

A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C. M'rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346) :174–180, May 2011.

[12] K. P. Aßhauer, B. Wemheuer, R. Daniel, and P. Meinicke. Tax4Fun : predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, 31(17) :2882–2884, Sept. 2015.

[13] F. Beghini, L. J. McIver, A. Blanco-Míguez, L. Dubois, F. Asnicar, S. Maharjan, A. Mailyan, P. Manghi, M. Scholz, A. M. Thomas, M. Valles-Colomer, G. Weingart, Y. Zhang, M. Zolfo, C. Huttenhower, E. A. Franzosa, and N. Segata. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10 :e65088, May 2021. Publisher : eLife Sciences Publications, Ltd.

[14] M. Bernard, O. Rué, M. Mariadassou, and G. Pascal. FROGS : a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. *Briefings in Bioinformatics*, 22(6) :bbab318, Nov. 2021.

[15] G. Biegert, M. B. El Alam, T. Karpinets, X. Wu, T. T. Sims, K. Yoshida-Court, E. J. Lynn, J. Yue, A. D. Medrano, J. Petrosino, M. P. Mezzari, N. J. Ajami, T. Solley, M. Ahmed-Kaddar, A. H. Klopp, and L. E. Colbert. Diversity and composition of gut microbiome of cervical cancer patients : Do results of 16S rRNA sequencing and whole genome sequencing approaches align ? *Journal of Microbiological Methods*, 185 :106213, June 2021.

[16] S. A. Boers, R. Jansen, and J. P. Hays. Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *European Journal of Clinical Microbiology & Infectious Diseases*, 38(6) :1059–1070, 2019.

[17] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15) :2114–2120, Aug. 2014.

[18] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J. E. Bisanz, K. Bittinger, A. Brejnrod, C. J. Brislawn, C. T. Brown, B. J. Callahan, A. M. Caraballo-Rodríguez, J. Chase, E. K. Cope, R. Da Silva, C. Diener, P. C. Dorrestein, G. M. Douglas, D. M. Durall, C. Duvallet, C. F. Edwardson, M. Ernst, M. Estaki, J. Fouquier, J. M. Gauglitz, S. M. Gibbons, D. L. Gibson, A. Gonzalez, K. Gorlick, J. Guo, B. Hillmann, S. Holmes, H. Holste, C. Huttenhower, G. A. Huttley, S. Janssen, A. K. Jarmusch, L. Jiang, B. D. Kaehler, K. B. Kang, C. R. Keefe, P. Keim, S. T. Kelley, D. Knights, I. Koester, T. Kosciolek, J. Kreps, M. G. I. Langille, J. Lee, R. Ley, Y.-X. Liu, E. Loftfield, C. Lozupone, M. Maher, C. Marotz, B. D. Martin, D. McDonald, L. J. McIver, A. V. Melnik, J. L. Metcalf, S. C. Morgan, J. T. Morton, A. T. Naimey, J. A. Navas-Molina, L. F. Nothias, S. B. Orchanian, T. Pearson, S. L. Peoples, D. Petras, M. L. Preuss, E. Pruesse, L. B. Rasmussen, A. Rivers, M. S. Robeson, P. Rosenthal, N. Segata, M. Shaffer, A. Shiffer, R. Sinha, S. J. Song, J. R. Spear, A. D. Swafford, L. R. Thompson, P. J. Torres, P. Trinh,

A. Tripathi, P. J. Turnbaugh, S. Ul-Hasan, J. J. J. van der Hooft, F. Vargas, Y. Vázquez-Baeza, E. Vogtmann, M. von Hippel, W. Walters, Y. Wan, M. Wang, J. Warren, K. C. Weber, C. H. D. Williamson, A. D. Willis, Z. Z. Xu, J. R. Zaneveld, Y. Zhang, Q. Zhu, R. Knight, and J. G. Caporaso. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8) :852–857, Aug. 2019. Number : 8 Publisher : Nature Publishing Group.

[19] F. Bonk, D. Popp, H. Harms, and F. Centler. PCR-based quantification of taxa-specific abundances in microbial communities : Quantifying and avoiding common pitfalls. *Journal of Microbiological Methods*, 153 :139–147, Oct. 2018.

[20] L. Breiman. Random Forests. *Machine Learning*, 45(1) :5–32, Oct. 2001.

[21] F. P. Breitwieser, J. Lu, and S. L. Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, 20(4) :1125–1136, Sept. 2017.

[22] M. S. O. Brieuc, C. D. Waters, D. P. Drinan, and K. A. Naish. A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, 18(4) :755–766, 2018. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12773.

[23] J. P. Brooks, G. A. Buck, G. Chen, L. Diao, D. J. Edwards, J. M. Fettweis, S. Huzurbazar, A. Rakitin, G. A. Satten, E. Smirnova, Z. Waks, M. L. Wright, C. Yanover, and Y.-H. Zhou. Changes in vaginal community state types reflect major shifts in the microbiome. *Microbial Ecology in Health and Disease*, 28(1) :1303265, Apr. 2017.

[24] C. S. Byrne, E. S. Chambers, D. J. Morrison, and G. Frost. The role of short chain fatty acids in appetite regulation and energy homeostasis. *International Journal of Obesity (2005)*, 39(9) :1331–1338, Sept. 2015.

[25] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. DADA2 : High resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7) :581–583, July 2016.

[26] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5) :335–336, May 2010.

[27] F. Cattonaro, A. Spadotto, S. Radovic, and F. Marroni. Do you cov me ? Effect of coverage reduction on metagenome shotgun sequencing studies. *F1000 Research*, 7 :1767, 2020.

[28] M. Chen, Y. Feng, and W. Liu. Efficacy and safety of probiotics in the induction and maintenance of inflammatory bowel disease remission : a systematic review and meta-analysis. *Annals of Palliative Medicine*, 10(11) :11821–11829, Nov. 2021.

[29] J. R. Cole, Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. Ribosomal Database Project : data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(Database issue) :D633–642, Jan. 2014.

[30] L. Coretti, L. Paparo, M. P. Riccio, F. Amato, M. Cuomo, A. Natale, L. Borrelli, G. Corrado, M. Comegna, E. Buommino, G. Castaldo, C. Bravaccio, L. Chiariotti, R. Berni Canani, and F. Lembo. Gut Microbiota Features in Young Children With Autism Spectrum Disorders. *Frontiers in Microbiology*, 9 :3146, 2018.

[31] A. Coscia, F. Bardanzellu, E. Caboni, V. Fanos, and D. G. Peroni. When a Neonate Is Born, So Is a Microbiota. *Life (Basel, Switzerland)*, 11(2) :148, Feb. 2021.

[32] P. I. Costea, F. Hildebrand, M. Arumugam, F. Bäckhed, M. J. Blaser, F. D. Bushman, W. M. de Vos, S. D. Ehrlich, C. M. Fraser, M. Hattori, C. Huttenhower, I. B. Jeffery, D. Knights, J. D. Lewis, R. E. Ley, H. Ochman, P. W. O'Toole, C. Quince, D. A. Relman, F. Shanahan, S. Sunagawa, J. Wang, G. M. Weinstock, G. D. Wu, G. Zeller, L. Zhao, J. Raes, R. Knight, and P. Bork. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3(1) :8–16, Jan. 2018.

[33] A. Cotillard, S. P. Kennedy, L. C. Kong, E. Prifti, N. Pons, E. Le Chatelier, M. Almeida, B. Quinquis, F. Levenez, N. Galleron, S. Gougis, S. Rizkalla, J.-M. Batto, P. Renault, J. Doré, J.-D. Zucker, K. Clément, and S. D. Ehrlich. Dietary intervention impact on gut microbial gene richness. *Nature*, 500(7464) :585–588, Aug. 2013. Number : 7464 Publisher : Nature Publishing Group.

[34] B. Cukrowska, J. B. Bierła, M. Zakrzewska, M. Klukowski, and E. Maciorkowska. The Relationship between the Infant Gut Microbiota and Allergy. The Role of Bifidobacterium breve and Prebiotic Oligosaccharides in the Activation of Anti-Allergic Mechanisms in Early Life. *Nutrients*, 12(4) :E946, Mar. 2020.

[35] L. P. da Silva, V. A. Mata, P. B. Lopes, R. J. Lopes, and P. Beja. High-resolution multi-marker DNA metabarcoding reveals sexual dietary differentiation in a bird with minor dimorphism. *Ecology and Evolution*, 10(19) :10364–10373, Sept. 2020.

[36] L. P. da Silva, V. A. Mata, P. B. Lopes, P. Pereira, S. N. Jarman, R. J. Lopes, and P. Beja. Advancing the integration of multi-marker metabarcoding data in dietary analysis of trophic generalists. *Molecular Ecology Resources*, 19(6) :1420–1432, Nov. 2019.

[37] D. Davar, A. K. Dzutsev, J. A. McCulloch, R. R. Rodrigues, J.-M. Chauvin, R. M. Morrison, R. N. Deblasio, C. Menna, Q. Ding, O. Pagliano, B. Zidi, S. Zhang, J. H. Badger, M. Vetizou, A. M. Cole, M. R. Fernandes, S. Prescott, R. G. F. Costa, A. K. Balaji, A. Morgun, I. Vujkovic-Cvijin, H. Wang, A. A. Borhani, M. B. Schwartz, H. M. Dubner, S. J. Ernst, A. Rose, Y. G. Najjar, Y. Belkaid, J. M. Kirkwood, G. Trinchieri, and H. M. Zarour. Fecal microbiota transplant overcomes resistance to anti–PD-1 therapy in melanoma patients. *Science*, 371(6529) :595–602, Feb. 2021. Publisher : American Association for the Advancement of Science Section : Research Article.

[38] G. A. de Groot, S. Geisen, E. R. J. Wubs, L. Meulenbroek, I. Laros, L. B. Snoek, D. R. Lammertsma, L. H. Hansen, and P. A. Slim. The aerobiome uncovered : Multi-marker metabarcoding reveals potential drivers of turn-over in the full microbial community in the air. *Environment International*, 154 :106551, Sept. 2021.

[39] F. De Luca and Y. Shoenfeld. The microbiome in autoimmune diseases. *Clinical and Experimental Immunology*, 195(1) :74–85, Jan. 2019.

[40] J. W. Debelius, M. Robeson, L. W. Hugerth, F. Boulund, W. Ye, and L. Engstrand. A comparison of approaches to scaffolding multiple regions along the 16S rRNA gene for improved resolution. preprint, Bioinformatics, Mar. 2021.

[41] K. Diop, J.-C. Dufour, A. Levasseur, and F. Fenollar. Exhaustive repertoire of human vaginal microbiota. *Human Microbiome Journal*, 11 :100051, Mar. 2019.

[42] G. M. Douglas, V. J. Maffei, J. R. Zaneveld, S. N. Yurgel, J. R. Brown, C. M. Taylor, C. Huttenhower, and M. G. I. Langille. PICRUSt2 for prediction of metagenome functions. *Nature biotechnology*, 38(6) :685–688, June 2020.

[43] F. Durazzi, C. Sala, G. Castellani, G. Manfreda, D. Remondini, and A. De Cesare. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, 11(1) :3030, Dec. 2021.

[44] F. Escudié, L. Auer, M. Bernard, M. Mariadassou, L. Cauquil, K. Vidal, S. Maman, G. Hernandez-Raquet, S. Combes, and G. Pascal. FROGS : Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, 34(8) :1287–1294, Apr. 2018.

[45] D. P. Faith, S. Veron, S. Pavoine, and R. Pellens. Indicators for the Expected Loss of Phylogenetic Diversity. In R. A. Scherson and D. P. Faith, editors, *Phylogenetic Diversity : Applications and Challenges in Biodiversity Science*, pages 73–91. Springer International Publishing, Cham, 2018.

[46] C. Fang, H. Zhong, Y. Lin, B. Chen, M. Han, H. Ren, H. Lu, J. M. Luber, M. Xia, W. Li, S. Stein, X. Xu, W. Zhang, R. Drmanac, J. Wang, H. Yang, L. Hammarström, A. D. Kostic, K. Kristiansen, and J. Li. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. *GigaScience*, 7(3) :1–8, Mar. 2018.

[47] V. Farrelly, F. A. Rainey, and E. Stackebrandt. Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Applied and Environmental Microbiology*, 61(7) :2798–2801, July 1995.

[48] Y. Feng, Y. Wang, B. Zhu, G. F. Gao, Y. Guo, and Y. Hu. Metagenome-assembled genomes and gene catalog from the chicken gut microbiome aid in deciphering antibiotic resistomes. *Communications Biology*, 4(1) :1–9, Nov. 2021. Number : 1 Publisher : Nature Publishing Group.

[49] A. D. Fernandes, J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor. Unifying the analysis of high-throughput sequencing datasets : characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2 :15, 2014.

[50] P. Ferretti, E. Pasolli, A. Tett, F. Asnicar, V. Gorfer, S. Fedi, F. Armanini, D. T. Truong, S. Manara, M. Zolfo, F. Beghini, R. Bertorelli, V. De Sanctis, I. Bariletti, R. Canto, R. Clementi, M. Cologna, T. Crifò, G. Cusumano, S. Gottardi, C. Innamorati, C. Masè, D. Postai, D. Savoi, S. Duranti, G. A. Lugli, L. Mancabelli, F. Turroni, C. Ferrario, C. Milani, M. Mangifesta, R. Anzalone, A. Viappiani, M. Yassour, H. Vlamakis, R. Xavier, C. M. Collado, O. Koren, S. Tateo, M. Soffiati, A. Pedrotti, M. Ventura, C. Huttenhower, P. Bork, and N. Segata. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host & Microbe*, 24(1) :133–145.e5, July 2018.

[51] A. A. Fodor, T. Z. DeSantis, K. M. Wylie, J. H. Badger, Y. Ye, T. Hepburn, P. Hu, E. Sodergren, K. Liolios, H. Huot-Creasy, B. W. Birren, and A. M. Earl. The "Most Wanted" Taxa from the Human Microbiome for Whole Genome Sequencing. *PLOS ONE*, 7(7) :e41294, 2012. Publisher : Public Library of Science.

[52] S. C. Forster, N. Kumar, B. O. Anonye, A. Almeida, E. Viciani, M. D. Stares, M. Dunn, T. T. Mkandawire, A. Zhu, Y. Shao, L. J. Pike, T. Louie, H. P. Browne, A. L. Mitchell, B. A. Neville, R. D. Finn, and T. D. Lawley. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology*, 37(2) :186–192, Feb. 2019.

[53] J. Frigerio, G. Agostinetto, V. Mezzasalma, F. De Mattia, M. Labra, and A. Bruno. DNA-Based Herbal Teas' Authentication : An ITS2 and psbA-trnH Multi-Marker DNA Metabarcoding Approach. *Plants*, 10(10) :2120, Oct. 2021.

[54] G. Fuks, M. Elgart, A. Amir, A. Zeisel, P. J. Turnbaugh, Y. Soen, and N. Shental. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome*, 6(1) :17, Dec. 2018.

[55] V. Gaboriau-Routhiau and N. Cerf-Bensussan. [Gut microbiota and development of the immune system]. *Medecine Sciences : M/S*, 32(11) :961–967, Nov. 2016.

[56] W. S. Garrett. Cancer and the microbiota. *Science (New York, N.Y.)*, 348(6230) :80–86, Apr. 2015.

[57] T. Gensollen, S. S. Iyer, D. L. Kasper, and R. S. Blumberg. How colonization by microbiota in early life shapes the immune system. *Science (New York, N.Y.)*, 352(6285) :539–544, Apr. 2016.

[58] J. S. Ghurye, V. Cepeda-Espinoza, and M. Pop. Metagenomic Assembly : Overview, Challenges and Applications. *The Yale Journal of Biology and Medicine*, 89(3) :353–362, Sept. 2016.

[59] J. Gilbert, M. J. Blaser, J. G. Caporaso, J. Jansson, S. V. Lynch, and R. Knight. Current understanding of the human microbiome. *Nature medicine*, 24(4) :392–400, Apr. 2018.

[60] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome Datasets Are Compositional : And This Is Not Optional. *Frontiers in Microbiology*, 8 :2224, Nov. 2017.

[61] G. B. Gloor and G. Reid. Compositional analysis : a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8) :692–703, Aug. 2016.

[62] S. Graspeuntner, N. Loeper, S. Künzel, J. F. Baines, and J. Rupp. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. *Scientific Reports*, 8(1) :9678, June 2018.

[63] M. Gurung, Z. Li, H. You, R. Rodrigues, D. B. Jump, A. Morgun, and N. Shulzhenko. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine*, 51 :102590, Jan. 2020.

[64] H. S. Gweon, L. P. Shaw, J. Swann, N. De Maio, M. AbuOun, R. Niehus, A. T. M. Hubbard, M. J. Bowes, M. J. Bailey, T. E. A. Peto, S. J. Hoosdally, A. S. Walker, R. P. Sebra, D. W. Crook, M. F. Anjum, D. S. Read, and N. Stoesser. The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environmental Microbiome*, 14(1) :7, Dec. 2019.

[65] C. Haifer, S. Paramsothy, N. O. Kaakoush, A. Saikal, S. Ghaly, T. Yang, L. D. W. Luu, T. J. Borody, and R. W. Leong. Lyophilised oral faecal microbiota transplantation for ulcerative colitis (LOTUS) : a randomised, double-blind, placebo-controlled trial. *The Lancet. Gastroenterology & Hepatology*, 7(2) :141–151, Feb. 2022.

[66] J. He, P. Zhang, L. Shen, L. Niu, Y. Tan, L. Chen, Y. Zhao, L. Bai, X. Hao, X. Li, S. Zhang, and L. Zhu. Short-Chain Fatty Acids and Their Association with Signalling Pathways in Inflammation, Glucose and Lipid Metabolism. *International Journal of Molecular Sciences*, 21(17) :E6356, Sept. 2020.

[67] A. D. Hershey and M. Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36(1) :39–56, May 1952.

[68] J. E. Hill, S. L. Penny, K. G. Crowell, S. H. Goh, and S. M. Hemmingsen. cpnDB : A Chaperonin Sequence Database. *Genome Research*, 14(8) :1669–1675, Aug. 2004.

[69] B. Hillmann, G. A. Al-Ghalith, R. R. Shields-Cutler, Q. Zhu, D. M. Gohl, K. B. Beckman, R. Knight, and D. Knights. Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems*, 3(6), Nov. 2018.

[70] C. Huang, P. Yi, M. Zhu, W. Zhou, B. Zhang, X. Yi, H. Long, G. Zhang, H. Wu, G. C. Tsokos, M. Zhao, and Q. Lu. Safety and efficacy of fecal microbiota transplantation for treatment of systemic lupus erythematosus : An EXPLORER trial. *Journal of Autoimmunity*, 130 :102844, June 2022.

[71] J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, and P. Bork. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8) :2115–2122, Aug. 2017.

[72] H. K. Hughes, D. Rose, and P. Ashwood. The Gut Microbiota and Dysbiosis in Autism Spectrum Disorders. *Current Neurology and Neuroscience Reports*, 18(11) :81, Sept. 2018.

[73] G. Ianiro, S. Bibbò, S. Porcari, C. R. Settanni, F. Giambò, A. R. Curta, G. Quaranta, F. Scaldaferri, L. Masucci, M. Sanguinetti, A. Gasbarrini, and G. Cammarota. Fecal microbiota transplantation for recurrent C. difficile infection in patients with inflammatory bowel disease : experience of a large-volume European FMT center. *Gut Microbes*, 13(1) :1994834, Dec. 2021.

[74] S. M. Jandhyala, R. Talukdar, C. Subramanyam, H. Vuyyuru, M. Sasikala, and D. N. Reddy. Role of the normal gut microbiota. *World Journal of Gastroenterology : WJG*, 21(29) :8787–8803, Aug. 2015.

[75] J. T. Jeske and C. Gallert. Microbiome Analysis via OTU and ASV-Based Pipelines-A Comparative Interpretation of Ecological Data in WWTP Systems. *Bioengineering (Basel, Switzerland)*, 9(4) :146, Mar. 2022.

[76] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter. InterProScan 5 : genome-scale protein function classification. *Bioinformatics (Oxford, England)*, 30(9) :1236–1240, May 2014.

[77] N. Kamada, G. Y. Chen, N. Inohara, and G. Núñez. Control of pathogens and pathobionts by the gut microbiota. *Nature Immunology*, 14(7) :685–690, July 2013.

[78] D. D. Kang, J. Froula, R. Egan, and Z. Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3 :e1165, 2015.

[79] E. Kartal, T. S. B. Schmidt, E. Molina-Montes, S. Rodríguez-Perales, J. Wirbel, O. M. Maistrenko, W. A. Akanni, B. Alashkar Alhamwe, R. J. Alves, A. Carrato, H.-P. Erasmus, L. Estudillo, F. Finkelmeier, A. Fullam, A. M. Glazek, P. Gómez-Rubio, R. Hercog,

F. Jung, S. Kandels, S. Kersting, M. Langheinrich, M. Márquez, X. Molero, A. Orakov, T. Van Rossum, R. Torres-Ruiz, A. Telzerow, K. Zych, MAGIC Study investigators, PanGenEU Study investigators, V. Benes, G. Zeller, J. Trebicka, F. X. Real, N. Malats, and P. Bork. A faecal microbiota signature with high specificity for pancreatic cancer. *Gut*, pages gutjnl–2021–324755, Mar. 2022.

[80] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14) :3059–3066, July 2002.

[81] S. W. Kembel, M. Wu, J. A. Eisen, and J. L. Green. Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Computational Biology*, 8(10) :e1002743, Oct. 2012.

[82] L. Khachatryan, R. H. de Leeuw, M. E. M. Kraakman, N. Pappas, M. Te Raa, H. Mei, P. de Knijff, and J. F. J. Laros. Taxonomic classification and abundance estimation using 16S and WGS-A comparison using controlled reference samples. *Forensic Science International. Genetics*, 46 :102257, May 2020.

[83] M. Kim, H.-S. Oh, S.-C. Park, and J. Chun. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 64(Pt_2) :346–351, Feb. 2014.

[84] Y. Kryachko, D. Semler, J. Vogrinetz, M. Lemke, R. Irvine, J. Davidson, M. G. Links, E. L. McCarthy, B. Haug, and S. M. Hemmingsen. Analyses of 16S rRNA and cpn60 gene sequences provide complementary information about potentially useful and harmful oil field microbiota. *International Biodeterioration & Biodegradation*, 123 :320–327, Sept. 2017.

[85] J.-C. Lagier, S. Khelaifia, M. T. Alou, S. Ndongo, N. Dione, P. Hugon, A. Caputo, F. Cadoret, S. I. Traore, E. H. Seck, G. Dubourg, G. Durand, G. Mourembou, E. Guilhot, A. Togo, S. Bellali, D. Bachar, N. Cassir, F. Bittar, J. Delerce, M. Mailhe, D. Ricaboni, M. Bilen, N. P. M. Dangui Nieko, N. M. Dia Badiane, C. Valles, D. Mouelhi, K. Diop, M. Million, D. Musso, J. Abrahão, E. I. Azhar, F. Bibi, M. Yasir, A. Diallo, C. Sokhna, F. Djossou, V. Vitton, C. Robert, J. M. Rolain, B. La Scola, P.-E. Fournier, A. Levasseur, and D. Raoult. Culture of previously uncultured members of the human gut microbiota by culturomics. *Nature Microbiology*, 1 :16203, Nov. 2016.

[86] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4) :357–359, Apr. 2012. Number : 4 Publisher : Nature Publishing Group.

[87] I. Laudadio, V. Fulci, F. Palone, L. Stronati, S. Cucchiara, and C. Carissimi. Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. *OMICS : A Journal of Integrative Biology*, 22(4) :248–254, Apr. 2018.

[88] E. Le Chatelier, Trine Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J.-M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jørgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clément, J. Doré, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J.-D. Zucker, J. Raes, T. Hansen, P. Bork, J. Wang, S. D. Ehrlich, and

O. Pedersen. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464) :541–546, Aug. 2013. Number : 7464 Publisher : Nature Publishing Group.

[89] S. J. Lee and M. Rho. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Scientific Reports*, 12(1) :824, Jan. 2022. Number : 1 Publisher : Nature Publishing Group.

[90] D. Li, R. Luo, C.-M. Liu, C.-M. Leung, H.-F. Ting, K. Sadakane, H. Yamashita, and T.-W. Lam. MEGAHIT v1.0 : A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102 :3–11, June 2016.

[91] H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv :1303.3997 [q-bio]*, May 2013. arXiv : 1303.3997.

[92] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, MetaHIT Consortium, P. Bork, J. Wang, and MetaHIT Consortium. An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, 32(8) :834–841, Aug. 2014.

[93] A. Limeta, B. Ji, M. Levin, F. Gatto, and J. Nielsen. Meta-analysis of the gut microbiota in predicting response to cancer immunotherapy in metastatic melanoma. *JCI insight*, 5(23) :140940, Dec. 2020.

[94] B.-N. Liu, X.-T. Liu, Z.-H. Liang, and J.-H. Wang. Gut microbiota in obesity. *World Journal of Gastroenterology*, 27(25) :3837–3850, July 2021.

[95] R. Loomba, V. Seguritan, W. Li, T. Long, N. Klitgord, A. Bhatt, P. S. Dulai, C. Caussy, R. Bettencourt, S. K. Highlander, M. B. Jones, C. B. Sirlin, B. Schnabl, L. Brinkac, N. Schork, C.-H. Chen, D. A. Brenner, W. Biggs, S. Yooseph, J. C. Venter, and K. E. Nelson. Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. *Cell Metabolism*, 25(5) :1054–1062.e5, May 2017.

[96] S. Louca, M. Doebeli, and L. W. Parfrey. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome*, 6(1) :41, Feb. 2018.

[97] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg. Bracken : estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3 :e104, Jan. 2017. Publisher : PeerJ Inc.

[98] J. Lu and S. L. Salzberg. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome*, 8(1) :124, Aug. 2020.

[99] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. Swarm v2 : highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3 :e1420, Dec. 2015.

[100] F. Malard, A. Vekhoff, S. Lapusan, F. Isnard, E. D'incan-Corda, J. Rey, C. Saillard, X. Thomas, S. Ducastelle-Lepretre, E. Paubelle, M.-V. Larcher, C. Rocher, C. Recher, S. Tavitian, S. Bertoli, A.-S. Michallet, L. Gilis, P. Peterlin, P. Chevallier, S. Nguyen, E. Plantamura, L. Boucinha, C. Gasc, M. Michallet, J. Dore, O. Legrand, and M. Mohty. Gut microbiota diversity after autologous fecal microbiota transfer in acute myeloid leukemia patients. *Nature Communications*, 12(1) :3084, May 2021.

[101] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, and S. D. Peddada. Analysis of composition of microbiomes : a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26 :27663, 2015.

[102] C. A. Marotz, J. G. Sanders, C. Zuniga, L. S. Zaramela, R. Knight, and K. Zengler. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*, 6(1) :42, Feb. 2018.

[103] C. Martin-Gallausiaux, L. Marinelli, H. M. Blottière, P. Larraufie, and N. Lapaque. SCFA : mechanisms and functional importance in the gut. *Proceedings of the Nutrition Society*, 80(1) :37–49, Feb. 2021. Publisher : Cambridge University Press.

[104] V. Matson, J. Fessler, R. Bao, T. Chongsuwat, Y. Zha, M.-L. Alegre, J. J. Luke, and T. F. Gajewski. The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science*, 359(6371) :104–108, Jan. 2018.

[105] J. A. McCulloch, D. Davar, R. R. Rodrigues, J. H. Badger, J. R. Fang, A. M. Cole, A. K. Balaji, M. Vetizou, S. M. Prescott, M. R. Fernandes, R. G. F. Costa, W. Yuan, R. Salcedo, E. Bahadiroglu, S. Roy, R. N. DeBlasio, R. M. Morrison, J.-M. Chauvin, Q. Ding, B. Zidi, A. Lowin, S. Chakka, W. Gao, O. Pagliano, S. J. Ernst, A. Rose, N. K. Newman, A. Morgun, H. M. Zarour, G. Trinchieri, and A. K. Dzutsev. Intestinal microbiota signatures of clinical response and immune-related adverse events in melanoma patients treated with anti-PD-1. *Nature Medicine*, 28(3) :545–556, Mar. 2022.

[106] A. J. McGuinness, J. A. Davis, S. L. Dawson, A. Loughman, F. Collier, M. O'Hely, C. A. Simpson, J. Green, W. Marx, C. Hair, G. Guest, M. Mohebbi, M. Berk, D. Stupart, D. Watters, and F. N. Jacka. A systematic review of gut microbiota composition in observational studies of major depressive disorder, bipolar disorder and schizophrenia. *Molecular Psychiatry*, 27(4) :1920–1935, Apr. 2022.

[107] P. J. McMurdie and S. Holmes. phyloseq : An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, 8(4) :e61217, Apr. 2013.

[108] P. Menzel, K. L. Ng, and A. Krogh. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7(1) :11257, Apr. 2016. Number : 1 Publisher : Nature Publishing Group.

[109] A. Milanese, D. R. Mende, L. Paoli, G. Salazar, H.-J. Ruscheweyh, M. Cuenca, P. Hingamp, R. Alves, P. I. Costea, L. P. Coelho, T. S. B. Schmidt, A. Almeida, A. L. Mitchell, R. D. Finn, J. Huerta-Cepas, P. Bork, G. Zeller, and S. Sunagawa. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications*, 10(1) :1014, Mar. 2019.

[110] C. Milani, G. Alessandri, M. Mangifesta, L. Mancabelli, G. A. Lugli, F. Fontana, G. Longhi, R. Anzalone, A. Viappiani, S. Duranti, F. Turroni, R. Costi, A. Annicchiarico, A. Morini, L. Sarli, M. C. Ossiprandi, D. van Sinderen, and M. Ventura. Untangling Species-Level Composition of Complex Bacterial Communities through a Novel Metagenomic Approach. *mSystems*, 5(4) :e00404–20, July 2020.

[111] C. Milani, S. Duranti, M. Mangifesta, G. A. Lugli, F. Turroni, L. Mancabelli, A. Viappiani, R. Anzalone, G. Alessandri, M. C. Ossiprandi, D. van Sinderen, and M. Ventura. Phylotype-Level Profiling of Lactobacilli in Highly Complex Environments by Means of an Internal Transcribed Spacer-Based Metagenomic Approach. *Applied and Environmental Microbiology*, 84(14), July 2018.

[112] P. Moayyedi, M. G. Surette, P. T. Kim, J. Libertucci, M. Wolfe, C. Onischi, D. Armstrong, J. K. Marshall, Z. Kassam, W. Reinisch, and C. H. Lee. Fecal Microbiota Transplantation Induces Remission in Patients With Active Ulcerative Colitis in a Randomized Controlled Trial. *Gastroenterology*, 149(1) :102–109.e6, July 2015.

[113] D. J. Morrison and T. Preston. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes*, 7(3) :189–200, May 2016.

[114] C. Mukherjee, C. J. Beall, A. L. Griffen, and E. J. Leys. High-resolution ISR amplicon sequencing reveals personalized oral microbiome. *Microbiome*, 6 :153, Sept. 2018.

[115] S. M. Naser, P. Dawyndt, B. Hoste, D. Gevers, K. Vandemeulebroecke, I. Cleenwerck, M. Vancanneyt, and J. Swings. Identification of lactobacilli by pheS and rpoA gene sequence analyses. *International Journal of Systematic and Evolutionary Microbiology*, 57(12) :2777–2789, Dec. 2007.

[116] S. M. Naser, F. L. Thompson, B. Hoste, D. Gevers, P. Dawyndt, M. Vancanneyt, and J. Swings. Application of multilocus sequence analysis (MLSA) for rapid identification of Enterococcus species based on rpoA and pheS genes. *Microbiology*, 151(7) :2141–2150, July 2005.

[117] S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, and N. C. Kyrpides. New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568(7753) :505–510, Apr. 2019.

[118] H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J.-M. Batto, M. B. Quintanilha Dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Doré, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Léonard, F. Levenez, O. Lund, B. Moumen, D. Le Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sørensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, MetaHIT Consortium, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, S. D. Ehrlich, and MetaHIT Consortium. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8) :822–828, Aug. 2014.

[119] V. L. Nikolova, M. R. B. Hall, L. J. Hall, A. J. Cleare, J. M. Stone, and A. H. Young. Perturbations in Gut Microbiota Composition in Psychiatric Disorders : A Review and Meta-analysis. *JAMA psychiatry*, 78(12) :1343–1354, Dec. 2021.

[120] R. H. Nilsson, K.-H. Larsson, A. F. Taylor, J. Bengtsson-Palme, T. S. Jeppesen, D. Schigel, P. Kennedy, K. Picard, F. O. Glöckner, L. Tedersoo, I. Saar, U. Kõljalg, and K. Abarenkov. The UNITE database for molecular identification of fungi : handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, 47(D1) :D259–D264, Jan. 2019.

[121] A. Nishida, R. Inoue, O. Inatomi, S. Bamba, Y. Naito, and A. Andoh. Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clinical Journal of Gastroenterology*, 11(1) :1–10, Feb. 2018.

[122] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner. metaSPAdes : a new versatile metagenomic assembler. *Genome Research*, 27(5) :824–834, May 2017.

[123] R. Nyanzi, P. J. Jooste, M. Cameron, and C. Witthuhn. Comparison of rpoA and pheS Gene Sequencing to 16S rRNA Gene Sequencing in Identification and Phylogenetic Analysis of LAB from Probiotic Food Products and Supplements. *Food Biotechnology*, 27(4) :303–327, Oct. 2013. Publisher : Taylor & Francis _eprint : https ://-doi.org/10.1080/08905436.2013.838783.

[124] M. A. Oberhardt, R. Zarecki, S. Gronow, E. Lang, H.-P. Klenk, U. Gophna, and E. Ruppin. Harnessing the landscape of microbial culture media to predict new organism–media pairings. *Nature Communications*, 6 :8493, Oct. 2015.

[125] J.-C. Ogier, S. Pagès, M. Galan, M. Barret, and S. Gaudriault. rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *BMC Microbiology*, 19(1) :1–16, Dec. 2019. Number : 1 Publisher : BioMed Central.

[126] T. G. Oh, S. M. Kim, T. Fu, J. Guo, S. Bassirian, S. Singh, E. V. Madamba, R. Bettencourt, L. Richards, M. Raffatellu, P. C. Dorrestein, R. T. Yu, A. R. Atkins, T. Huan, D. A. Brenner, C. B. Sirlin, R. Knight, M. Downes, R. M. Evans, and R. Loomba. A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. *Cell Metabolism*, page S1550413120303065, June 2020.

[127] C. V. Olovo, X. Huang, X. Zheng, and M. Xu. Faecal microbial biomarkers in early diagnosis of colorectal cancer. *Journal of Cellular and Molecular Medicine*, 25(23) :10783–10797, Dec. 2021.

[128] B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1) :385, Sept. 2011.

[129] A. E. Parada, D. M. Needham, and J. A. Fuhrman. Every base matters : assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, 18(5) :1403–1414, 2016. _eprint : https ://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.13023.

[130] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. CheckM : assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7) :1043–1055, July 2015.

[131] V. Pascal, M. Pozuelo, N. Borruel, F. Casellas, D. Campos, A. Santiago, X. Martinez, E. Varela, G. Sarrabayrouse, K. Machiels, S. Vermeire, H. Sokol, F. Guarner, and C. Manichanh. A microbial signature for Crohn's disease. *Gut*, 66(5) :813–822, May 2017.

[132] E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, and N. Segata. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176(3) :649–662.e20, Jan. 2019.

[133] E. Pasolli, L. Schiffer, P. Manghi, A. Renson, V. Obenchain, D. T. Truong, F. Beghini, F. Malik, M. Ramos, J. B. Dowd, C. Huttenhower, M. Morgan, N. Segata, and L. Waldron. Accessible, curated metagenomic data through ExperimentHub. *Nature methods*, 14(11) :1023–1024, Oct. 2017.

[134] N. K. Pinna, A. Dutta, M. Monzoorul Haque, and S. S. Mande. Can Targeting Non-Contiguous V-Regions With Paired-End Sequencing Improve 16S rRNA-Based Taxonomic Resolution of Microbiomes ? : An In Silico Evaluation. *Frontiers in Genetics*, 10 :653, July 2019.

[135] F. Plaza Oñate, E. Le Chatelier, M. Almeida, A. C. L. Cervino, F. Gauthier, F. Magoulès, S. D. Ehrlich, and M. Pichaud. MSPminer : abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics (Oxford, England)*, 35(9) :1544–1552, May 2019.

[136] M. Pohar, M. Blas, and S. Turk. Comparison of logistic regression and linear discriminant analysis : a simulation study. *Metodoloski zvezki*, 1(1) :143, 2004. ISBN : 1854-0023 Publisher : Anuska Ferligoj.

[137] S. Poirier, O. Rué, R. Peguilhan, G. Coeuret, M. Zagorec, M.-C. Champomier-Vergès, V. Loux, and S. Chaillou. Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing : A comparative analysis with 16S rDNA V3-V4 amplicon sequencing. *PLoS ONE*, 13(9) :e0204629, Sept. 2018.

[138] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3) :e9490, Mar. 2010.

[139] A. Prodan, V. Tremaroli, H. Brolin, A. H. Zwinderman, M. Nieuwdorp, and E. Levin. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PloS One*, 15(1) :e0227434, 2020.

[140] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang. A human gut microbial gene catalog established by metagenomic sequencing. *Nature*, 464(7285) :59–65, Mar. 2010.

[141] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. L. Chatelier, J. Yao, L. Wu, J. Zhou, S. Ni, L. Liu, N. Pons, J. M. Batto, S. P. Kennedy, P. Leonard, C. Yuan, W. Ding, Y. Chen, X. Hu, B. Zheng, G. Qian, W. Xu, S. D. Ehrlich, S. Zheng, and L. Li. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516) :59–64, Sept. 2014. Number : 7516 Publisher : Nature Publishing Group.

[142] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. The SILVA ribosomal RNA gene database project : improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue) :D590–D596, Jan. 2013.

[143] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9) :833–844, Sept. 2017.

[144] J. Ravel, P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, R. M. Brotman, C. C. Davis, K. Ault, L. Peralta, and L. J. Forney. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences of the United States of America*, 108(Suppl 1) :4680–4687, Mar. 2011.

[145] E. Rinninella, P. Raoul, M. Cintoni, F. Franceschi, G. A. D. Miggiano, A. Gasbarrini, and M. C. Mele. What is the Healthy Gut Microbiota Composition ? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms*, 7(1) :E14, Jan. 2019.

[146] R. C. Robertson, A. R. Manges, B. B. Finlay, and A. J. Prendergast. The Human Microbiome and Child Growth – First 1000 Days and Beyond. *Trends in Microbiology*, 27(2) :131–147, Feb. 2019.

[147] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. VSEARCH : a versatile open source tool for metagenomics. *PeerJ*, 4 :e2584, 2016.

[148] H. Roume, E. Le Chaterlier, N. Pons, and D. Ehrlich. Impact of decreasing the sequencing depth onto gut microbiota analysis using shotgun metagenomics approach. *Poster at the 7th International Human Microbiome Consortium Meeting, Killarney, Ireland*, June 2018.

[149] E. Rutayisire, K. Huang, Y. Liu, and F. Tao. The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life : a systematic review. *BMC gastroenterology*, 16(1) :86, July 2016.

[150] T. M. Santiago-Rodriguez, A. Garoutte, E. Adams, W. Nasser, M. C. Ross, A. La Reau, Z. Henseler, T. Ward, D. Knights, J. F. Petrosino, and E. B. Hollister. Metagenomic Information Recovery from Human Stool Samples Is Influenced by Sequencing Depth and Profiling Method. *Genes*, 11(11) :E1380, Nov. 2020.

[151] E. Saus, S. Iraola-Guzmán, J. R. Willis, A. Brunet-Vega, and T. Gabaldón. Microbiome and colorectal cancer : Roles in carcinogenesis and clinical potential. *Molecular Aspects of Medicine*, 69 :93–106, Oct. 2019.

[152] J. Schellenberg, M. G. Links, J. E. Hill, T. J. Dumonceaux, G. A. Peters, S. Tyler, T. B. Ball, A. Severini, and F. A. Plummer. Pyrosequencing of the Chaperonin-60 Universal Target as a Tool for Determining Microbial Community Composition. *Applied and Environmental Microbiology*, 75(9) :2889–2898, May 2009.

[153] K. H. Schleifer. Classification of Bacteria and Archaea : past, present and future. *Systematic and Applied Microbiology*, 32(8) :533–542, Dec. 2009.

[154] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. Introducing mothur : Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23) :7537–7541, Dec. 2009. Publisher : American Society for Microbiology.

[155] A. E. Schriefer, P. F. Cliften, M. C. Hibberd, C. Sawyer, V. Brown-Kennerly, L. Burcea, E. Klotz, S. D. Crosby, J. I. Gordon, and R. D. Head. A multi-amplicon 16S rRNA sequencing and analysis method for improved taxonomic profiling of bacterial communities. *Journal of Microbiological Methods*, 154 :6–13, Nov. 2018.

[156] T. Seemann. Prokka : rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14) :2068–2069, July 2014.

[157] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8) :811–814, Aug. 2012. Number : 8 Publisher : Nature Publishing Group.

[158] R. Sender, S. Fuchs, and R. Milo. Are We Really Vastly Outnumbered ? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, 164(3) :337–340, Jan. 2016.

[159] F. Shanahan, T. S. Ghosh, and P. W. O'Toole. The Healthy Microbiome-What Is the Definition of a Healthy Gut Microbiome ? *Gastroenterology*, 160(2) :483–494, Jan. 2021.

[160] Y. Shi, G. Wang, H. C.-H. Lau, and J. Yu. Metagenomic Sequencing for Microbial DNA in Human Samples : Emerging Technological Advances. *International Journal of Molecular Sciences*, 23(4) :2181, Feb. 2022.

[161] S. Shirvani-Rad, O. Tabatabaei-Malazy, S. Mohseni, S. Hasani-Ranjbar, A.-R. Soroush, Z. Hoseini-Tavassol, H.-S. Ejtahed, and B. Larijani. Probiotics as a Complementary Therapy for Management of Obesity : A Systematic Review. *Evidence-Based Complementary and Alternative Medicine : eCAM*, 2021 :6688450, 2021.

[162] S. Silvaraju, N. Menon, H. Fan, K. Lim, and S. Kittelmann. Phylotype-Level Characterization of Complex Communities of Lactobacilli Using a High-Throughput, High-Resolution Phenylalanyl-tRNA Synthetase (*pheS*) Gene Amplicon Sequencing Approach. *Applied and Environmental Microbiology*, 87(1), Dec. 2020.

[163] T. G. Simon, A. T. Chan, and C. Huttenhower. Microbiome Biomarkers : One Step Closer in NAFLD Cirrhosis. *Hepatology (Baltimore, Md.)*, 73(5) :2063–2066, May 2021.

[164] A. Sirichoat, N. Sankuntaw, C. Engchanil, P. Buppasiri, K. Faksri, W. Namwat, W. Chantratita, and V. Lulitanond. Comparison of different hypervariable regions of 16S rRNA for taxonomic profiling of vaginal microbiota using next-generation sequencing. *Archives of Microbiology*, 203(3) :1159–1166, Apr. 2021.

[165] A. Sivan, L. Corrales, N. Hubert, J. B. Williams, K. Aquino-Michaels, Z. M. Earley, F. W. Benyamin, Y. M. Lei, B. Jabri, M.-L. Alegre, E. B. Chang, and T. F. Gajewski. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science (New York, N.Y.)*, 350(6264) :1084–1089, Nov. 2015.

[166] H. Sokol, C. Landman, P. Seksik, L. Berard, M. Montil, I. Nion-Larmurier, A. Bourrier, G. Le Gall, V. Lalande, A. De Rougemont, J. Kirchgesner, A. Daguenel, M. Cachanado, A. Rousseau, Drouet, M. Rosenzwajg, H. Hagege, X. Dray, D. Klatzman, P. Marteau, Saint-Antoine IBD Network, L. Beaugerie, and T. Simon. Fecal microbiota transplantation to maintain remission in Crohn's disease : a pilot randomized controlled study. *Microbiome*, 8(1) :12, Feb. 2020.

[167] R. Starke, V. S. Pylro, and D. K. Morais. 16S rRNA Gene Copy Number Normalization Does Not Provide More Reliable Conclusions in Metataxonomic Surveys. *Microbial Ecology*, 81(2) :535–539, Feb. 2021.

[168] S. Stefanni, D. Stanković, D. Borme, A. de Olazabal, T. Juretić, A. Pallavicini, and V. Tirelli. Multi-marker metabarcoding approach to study mesozooplankton at basin scale. *Scientific Reports*, 8 :12085, Aug. 2018.

[169] S. F. Stoddard, B. J. Smith, R. Hein, B. R. Roller, and T. M. Schmidt. rrnDB : improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*, 43(Database issue) :D593–D598, Jan. 2015.

[170] S. Sun, R. B. Jones, and A. A. Fodor. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome*, 8(1) :46, Apr. 2020.

[171] S. Tamburini, N. Shen, H. C. Wu, and J. C. Clemente. The microbiome in early life : implications for health outcomes. *Nature Medicine*, 22(7) :713–722, July 2016.

[172] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402) :207–214, June 2012.

[173] H. Tilg, P. D. Cani, and E. A. Mayer. Gut microbiome and liver diseases. *Gut*, 65(12) :2035–2044, Dec. 2016.

[174] L. Topstad, R. Guidetti, M. Majaneva, and T. Ekrem. Multi-marker DNA metabarcoding reflects tardigrade diversity in different habitats. *Genome*, 64(3) :217–231, Mar. 2021.

[175] M. L. Treiber, D. H. Taft, I. Korf, D. A. Mills, and D. G. Lemay. Pre- and post-sequencing recommendations for functional annotation of human fecal metagenomes. *BMC Bioinformatics*, 21(1) :74, Dec. 2020.

[176] D. Tz, H. P, L. N, R. M, B. El, K. K, H. T, D. D, H. P, and A. Gl. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, 72(7), July 2006. Publisher : Appl Environ Microbiol.

[177] L. W. J. van den Elsen, J. Garssen, R. Burcelin, and V. Verhasselt. Shaping the Gut Microbiota by Breastfeeding : The Gateway to Allergy Prevention ? *Frontiers in Pediatrics*, 7 :47, 2019.

[178] W. J. Van Der Pol, R. Kumar, C. D. Morrow, E. E. Blanchard, C. M. Taylor, D. H. Martin, E. J. Lefkowitz, and C. A. Muzny. In Silico and Experimental Evaluation of Primer Sets for Species-Level Resolution of the Vaginal Microbiota Using 16S Ribosomal RNA Gene Sequencing. *The Journal of Infectious Diseases*, 219(2) :305–314, Jan. 2019.

[179] S. J. Vancuren and J. E. Hill. Update on cpnDB : a reference database of chaperonin sequences. *Database : The Journal of Biological Databases and Curation*, 2019 :baz033, Mar. 2019.

[180] S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PloS One*, 7(8) :e42671, 2012.

[181] R. Villéger, A. Lopès, J. Veziant, J. Gagnière, N. Barnich, E. Billard, D. Boucher, and M. Bonnet. Microbial markers in colorectal cancer detection and/or prognosis. *World Journal of Gastroenterology*, 24(22) :2327–2347, June 2018.

[182] M. Vos, C. Quince, A. S. Pijl, M. d. Hollander, and G. A. Kowalchuk. A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity. *PLOS ONE*, 7(2) :e30600, 2012. Publisher : Public Library of Science.

[183] T. Větrovský and P. Baldrian. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE*, 8(2) :e57923, Feb. 2013.

[184] L. Wampach, A. Heintz-Buschart, J. V. Fritz, J. Ramiro-Garcia, J. Habier, M. Herold, S. Narayanasamy, A. Kaysen, A. H. Hogan, L. Bindl, J. Bottu, R. Halder, C. Sjöqvist, P. May, A. F. Andersson, C. de Beaufort, and P. Wilmes. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nature Communications*, 9 :5091, Nov. 2018.

[185] S. Wang, B. Sun, J. Tu, and Z. Lu. Improving the microbial community reconstruction at the genus level by multiple 16S rRNA regions. *Journal of Theoretical Biology*, 398 :1–8, June 2016.

[186] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids : A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356) :737–738, Apr. 1953. Number : 4356 Publisher : Nature Publishing Group.

[187] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain : the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11) :5088–5090, Nov. 1977.

[188] S. H. Wong and J. Yu. Gut microbiota in colorectal cancer : mechanisms of action and clinical applications. *Nature Reviews Gastroenterology & Hepatology*, 16(11) :690–704, Nov. 2019.

[189] D. E. Wood, J. Lu, and B. Langmead. Improved metagenomic analysis with Kraken 2. preprint, Bioinformatics, Sept. 2019.

[190] L. Xiao, J. Estellé, P. Kiilerich, Y. Ramayo-Caldas, Z. Xia, Q. Feng, S. Liang, A. Pedersen, N. J. Kjeldsen, C. Liu, E. Maguin, J. Doré, N. Pons, E. Le Chatelier, E. Prifti, J. Li, H. Jia, X. Liu, X. Xu, S. D. Ehrlich, L. Madsen, K. Kristiansen, C. Rogel-Gaillard, and J. Wang. A reference gene catalogue of the pig gut microbiome. *Nature Microbiology*, 1(12) :1–6, Sept. 2016. Number : 12 Publisher : Nature Publishing Group.

[191] F. Xie, W. Jin, H. Si, Y. Yuan, Y. Tao, J. Liu, X. Wang, C. Yang, Q. Li, X. Yan, L. Lin, Q. Jiang, L. Zhang, C. Guo, C. Greening, R. Heller, L. L. Guan, P. B. Pope, Z. Tan, W. Zhu, M. Wang, Q. Qiu, Z. Li, and S. Mao. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome*, 9(1) :137, June 2021.

[192] H. Xie, C. Yang, Y. Sun, Y. Igarashi, T. Jin, and F. Luo. PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning. *Frontiers in Genetics*, 11 :516269, 2020.

[193] M. Xie, M. Pan, Y. Jiang, X. Liu, W. Lu, J. Zhao, H. Zhang, and W. Chen. groEL Gene-Based Phylogenetic Analysis of Lactobacillus Species by High-Throughput Sequencing. *Genes*, 10(7) :530, July 2019.

[194] A. York. Microbiome : Gut microbiota sways response to cancer immunotherapy. *Nature Reviews. Microbiology*, 16(3) :121, Mar. 2018.

[195] R. Zaheer, N. Noyes, R. Ortega Polo, S. R. Cook, E. Marinier, G. Van Domselaar, K. E. Belk, P. S. Morley, and T. A. McAllister. Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*, 8(1) :5890, Dec. 2018.

[196] D. Zeevi, T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, J. Suez, J. A. Mahdi, E. Matot, G. Malka, N. Kosower, M. Rein, G. Zilberman-Schapira, L. Dohnalová, M. Pevsner-Fischer, R. Bikovsky, Z. Halpern, E. Elinav, and E. Segal. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*, 163(5) :1079–1094, Nov. 2015. Publisher : Elsevier.

[197] C.-B. Zhou, Y.-L. Zhou, and J.-Y. Fang. Gut Microbiota in Cancer Immune Response and Immunotherapy. *Trends in Cancer*, 7(7) :647–660, July 2021.

[198] M. Ziemski, T. Wisanwanichthan, N. A. Bokulich, and B. D. Kaehler. Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences. *Frontiers in Microbiology*, 12 :644487, 2021.

[199] Y. Zou, W. Xue, G. Luo, Z. Deng, P. Qin, R. Guo, H. Sun, Y. Xia, S. Liang, Y. Dai, D. Wan, R. Jiang, L. Su, Q. Feng, Z. Jie, T. Guo, Z. Xia, C. Liu, J. Yu, Y. Lin, S. Tang, G. Huo, X. Xu, Y. Hou, X. Liu, J. Wang, H. Yang, K. Kristiansen, J. Li, H. Jia, and

L. Xiao. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, 37(2) :179–185, Feb. 2019.

[200] V. Álvarez Arraño and S. Martín-Peláez. Effects of Probiotics and Synbiotics on Weight Loss in Subjects with Overweight or Obesity : A Systematic Review. *Nutrients*, 13(10) :3627, Oct. 2021.